Karlsruher Institut für Technologie

Masterarbeit

Vergleich von statistischen Lernverfahren und eine Anwendung in der Pneumologie

Julian Frank

29. Oktober 2015

Betreuung: Dr. Bernhard Klar

Fakultät für Mathematik

Karlsruher Institut für Technologie

Inhaltsverzeichnis

1 Einleitung

Der Karlsruher Facharzt für innere Medizin und Pneumologie Herr Dr. med. Johannes Schildge hat von 806 seiner an interstitiellen Lungenerkrankungen (ILD) leidenden Patienten Daten erhoben. Neben der obligatorischen Erfassung der Lungenfunktionswerte, epidemiologischer Größen wie Geschlecht, Alter und Angaben zum Rauchverhalten, werden vorallem Informationen aus der Auswertung einer bestimmten Lungenspülung, der bronchoalveolären Lavage (BAL), gewonnen. Die Zusammenhänge zwischen diesen Merkmalen und den interstitiellen Lungenkrankheiten wurden bereits in medizinischen Studien statistisch untersucht. Kapitel 6.3 fasst die relevantesten Ergebnisse und die unterschiedlichen Charakteristiken der Krankheiten zusammen, während in Kapitel 6.1 und 6.2 alle weiteren medizinischen Fachbegriffe definiert werden. Das wissenschaftliche Novum der Untersuchungen von Herrn Dr. Schildge ist die Bestimmung des Proteingehalts in der BAL. Ziel dieser Arbeit ist es, den Zusammenhang zwischen Protein und ILDs statistisch zu untersuchen, sowie die angepassten Modelle quantitativ und die dafür verwendeten Methoden theoretisch miteinander zu vergleichen.
Eine zweifelsfreie Diagnose an welcher ILD ein Patient leidet, kann oftmals auch mit radiologischem Befund und nach erfolgter Lungenbiopsie nicht getroffen werden. Es stellt sich aus mathematischer Sicht die Aufgabe, eine Prognose zu treffen, an welcher Krankheit ein Patient am wahrscheinlichsten leidet. Eine klassische statistische Modellierung solcher Probleme, die Bayessche Entscheidungstheorie, wird in Kapitel 2 skizziert.
Die entsprechenden Schätzer heißen Klassifikatoren und ordnen einer Beobachtung eine Kategorie zu. Wie lassen sich Begriffe wie Varianz und Bias für solche Schätzer kategorialer Merkmale verallgemeinern? Antworten findet man in den Arbeiten von Forschern auf dem Gebiet des maschinellen Lernens. Zwei relevante Veröffentlichungen werden hierfür in Kapitel 3.2 diskutiert und im Rahmen eines Reviews unter neuen Aspekten beleuchtet.
Das statistische Werkzeug, Methoden, die durch Beobachtung lernen, Daten zu klassifizieren, werden in Kapitel 4 ausführlich vorgestellt. Resampling-Methoden wie Kreuzvalidierung, Bootstrapping und Bagging verwenden zufällige Teilmengen des Datensatzes zur Schätzung der Fehlklassifikationen und zur Verbesserung des Klassifikators (s. Kapitel 3.3 und 5.1). Die als Meilensteine des maschinellen Lernens betrachteten Boosting-Algorithmen ermöglichen es einer Klassifikationsmethode schrittweise aus ihren Fehlern zu lernen.
In Kapitel 7 werden die von Herrn Dr. Schildge erfassten Daten durch Zufallsvariablen statistisch modelliert und Aussagen über ihre Verteilungen getroffen. Durch die Beta-Binomialverteilung wird eine Parametrisierung für die Verteilung der Anzahl der Immunzellen in der BAL vorgeschlagen. Eine Klassifizierung der Patienten in die verschiedenen Krankheitsgruppen erfolgt in Kapitel 8. Die zentrale Leitfrage ist dabei, ob und wie gut sich das Protein in der BAL als Prognosevariable eignet. Kapitel 9 beantwortet zudem, ob sich durch Proteinwerte eine idiopathische Lungenfibrose, eine ILD mit hoher Mortalität, ausschließen lässt.

2 Bayessche Entscheidungstheorie

Zufällige Beobachtungen kategorialer Daten werden durch eine Zufallsvariable G modelliert, die Werte in einer diskreten Menge $\mathcal{G}$ annimmt. Sei dabei $\mathcal{G} = \{1, \ldots, K\}$, wobei jede der K Ziffern für eine Gruppe steht.
Unterzieht sich ein Patient einer medizinischen Untersuchung, so gehört er entweder zur Gruppe der Gesunden oder leidet unter einer von $K-1$ Krankheiten. Solange ein Arzt nichts über seinen Patienten weiß, kann er lediglich die Aussage treffen, dass er mit einer *A-priori-Wahrscheinlichkeit* von

$$\pi_k = \mathbb{P}\left(G = k\right)$$

zur k-ten Gruppe gehört; die Wahrscheinlichkeit, mit der ein jeder Mensch an dieser Krankheit leidet, bzw. gesund ist. Im Rahmen einer Anamnese und medizinischer Tests werden Informationen über Symptome und Risikofaktoren qualitativer Art, sowie organische Funktionswerte quantitativer Art gewonnen, die durch einen Zufallsvektor X modelliert werden können. Mit Hilfe der beobachteten Daten x und den *A-posteriori-Wahrscheinlichkeiten*

$$p_k(x) = \mathbb{P}\left(G = k \mid X = x\right),\ k = 1, \ldots, K$$

kann man bestimmte Gruppen ausschließen, um so zu einer Diagnose zu gelangen. Diese muss nicht immer hundertprozentig korrekt sein, z. B. durch falsch positive Ergebnisse oder schwer zu unterscheidende Krankheitsbilder. Gegebenenfalls sind noch weitere Untersuchungen notwendig.
Die abhängige Variable G sei im Folgenden auch als Ausgabewert bezeichnet, die unabhängige X als Eingabewert, die Realisierungen in einem Eingaberaum $\mathcal{X}$ annimmt. Dabei sei für diese Arbeit $\mathcal{X}$ der p-dimensionale reelle Vektorraum.
Das Ziel von überwachten Klassifikationsverfahren ist es, auf Grundlage von N beobachteten Ein- und Ausgabewerten die Gruppenzugehörigkeit einer neuen Beobachtung x möglichst genau vorherzusagen. Der i-te Eingabewert sei mit x_i bezeichnet und der i-te Ausgabewert als g_i. Die Gesamtheit aller Daten $(x_i, g_i),\ i = 1, \ldots, N$ heißt in diesem Kontext *Trainingsdatensatz*. Ein Schätzer $G(x)$, der einer neuen Beobachtung einen Wert in $\mathcal{G}$ zuordnet, heißt *Klassifikator* oder *Klassifikations-/Entscheidungsregel*. Sind die A-posteriori-Wahrscheinlichkeiten bekannt, so ist der *Bayes-Klassifikator* definiert als

$$G^*(x) = \arg\max_{k \in \mathcal{G}} p_k(x). \tag{2.1}$$

Da die zugrundeliegenden Verteilungen in der Realität nicht bekannt sind, ist es ein naheliegender Ansatz, die Wahrscheinlichkeiten $p_k(x;\vartheta) = \mathbb{P}_\vartheta(G = k \mid X = x)$ zu parametrisieren um durch einen Schätzer $\hat{\vartheta}$ den Klassifikator

$$\hat{G}(x) = G(x;\hat{\vartheta}) = \arg\max_{k \in \mathcal{G}} p_k(x;\hat{\vartheta}) \tag{2.2}$$

zu erhalten. Alternativ führt die Modellierung einer streng monotonen Transformation, z. B. $f_k(x;\vartheta) = \log p_k(x;\vartheta)$ auf die Klassifikationsregel

$$\hat{G}(x) = G(x;\hat{\vartheta}) = \arg\max_{k \in \mathcal{G}} f_k(x;\hat{\vartheta}). \tag{2.3}$$

Dieses Vorgehen findet sich in ähnlicher Weise bei der linearen Diskriminanzanalyse (s. Kapitel 4.1) und der multinomialen logistischen Regression (s. Kapitel 4.2) wieder. Diese Grundbegriffe sind aus dem Standardlehrwerk *The Elements of Statistical Learning* [7, S. 9-11 u. S. 21] von Hastie, Tibshirani und Friedman entnommen.

3 Modellselektion und -bewertung

3.1 Verlustfunktion

Durch eine Verlustfunktion $L : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_{\geq 0},\ (k, \hat{k}) \mapsto L(k, \hat{k})$ kann ein Ausgabewert mit seiner Schätzung verglichen werden, um die Qualität der Prognose einer Klassifikationsregel zu beurteilen. Da die Menge, in der die abhängige Variable Werte annimmt, diskret ist, liegt die Wahl der zur diskreten Metrik analogen *0/1-Verlustfunktion* nahe.

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$$

Ein zufälliger Trainingsdatensatz ist mit den unabhängig identisch verteilten Zufallsvektoren $(X_1, G_1), \ldots, (X_N, G_N) \sim (X, G)$ durch $Z = ((X_1, G_1), \ldots, (X_N, G_N))$ beschrieben. Er ist dabei insbesondere unabhängig von einer zukünftigen Beobachtung (X, G). Passt man nun ein Modell zur Klassifikation an eine Realisierung z an, so erhält man eine Klassifikationsregel, die im Folgenden mit $\hat{G}(x, z)$ notiert wird. In diesem Zusammenhang heißt $\hat{G}$ *Klassifikationsmethode* und $\hat{G}(\cdot, z)$ bezeichnet die an den Datensatz z angepasste Entscheidungsregel der Methode. Dabei wird die Notation von $\hat{G}$ um das Argument z erweitert, damit klar ist, ob der Trainingsdatensatz im jeweiligen Kontext als Zufallsvektor oder als Realisation zu verstehen ist. Der *Testfehler*

$$\mathrm{Err}_z(\hat{G}) = \mathbb{E}L(G, \hat{G}(X, z)) = \mathbb{P}(G \neq \hat{G}(X, z))$$

beschreibt die Wahrscheinlichkeit, dass die Klassifikationsregel einen zukünftigen Eingabewert fehlklassifiziert. Er kann durch den *Trainingsfehler*

$$\mathrm{err}(\hat{G}) = \frac{1}{N} \sum_{i=1}^{N} L(g_i, \hat{G}(x_i, z)) = \frac{1}{N} \sum_{i=1}^{N} I(g_i \neq \hat{G}(x_i, z))$$

geschätzt werden. Je komplexer ein Modell gewählt wird, z. B. durch eine immer größere Anzahl an zu schätzenden Parametern, desto mehr kann es an den Trainingsdatensatz angepasst werden, was einen geringeren Testfehler zur Folge hat. Jedoch steigt mit der Komplexität unter Umständen auch die Varianz, da kleine Veränderungen des Trainingsdatensatzes die Klassifikationsregel entscheidend verändern und damit die Zuverlässigkeit der Vorhersage negativ beeinflussen. Vorschläge für eine mathematische Definition von Varianz eines Klassifikators werden in Kapitel 3.2 erläutert. Für eine Varianz, die die Variabilität einer Klassifikationsmethode, verursacht durch die Zufälligkeit des Trainingsdatensatzes, beschreibt, schlägt Domingos [6] entsprechende Definitionen

vor. Daher ist unter der Berücksichtigung, dass der Trainingsdatensatz selbst zufällig ist, der *erwartete Testfehler*

$$\mathrm{Err}(\hat{G}) = \mathbb{E}L(G, \hat{G}(X, Z)) = \mathbb{E}\mathrm{Err}_Z(\hat{G})$$

ein Maß für die zu erwartende Fehlklassifikation der Methode. Hastie, Tibshirani und Friedman [7, S. 219-221] unterscheiden dadurch klar zwischen dem Fehler einer Klassifikationsregel (Testfehler) und dem einer Klassifikationsmethode (erwarteter Testfehler).

3.2 Varianz und Bias von Klassifikatoren

Die Varianz kann für einen Klassifikator nicht wie bei einer reellen Zufallsvariable definiert werden, da seine Natur eine nominale ist. Die hier vorgestellten Theorien von Varianz und Bias von kategorialen Zufallsvariablen sind nicht ausgereift und daher nicht Teil von Lehrbüchern des maschinellen Lernens. Zu den vorgeschlagenen deutschen Übersetzungen der Fachbegriffe sind jeweils die Originaldefinitionen auf Englisch angegeben. Zudem wird für dieses Kapitel eine eigene, stellenweise ausführlichere Notation verwendet.
James [8] diskutiert verschiedene Vorschläge zur Definition von Varianz und Bias anderer Autoren und entwirft seinen eigenen Formalismus für eine Bias-Varianz-Zerlegung für beliebige symmetrische Verlustfunktionen. Seine Ausführungen werden in diesem Kapitel kurz zusammengefasst und seine Anmerkung zu den Definitionen von Domingos [6] näher erläutert.
Zunächst werden die Begriffe für den bekannten Fall eines reellen Ausgabewertes Y wiederholt, um die Verallgemeinerungen zu motivieren. Sei dafür $\hat{Y} = \hat{Y}(X)$ ein an den Datensatz z angepasstes Regressionsmodell mit dem Eingabevektor, bzw. mit den Regressoren X. Die Varianz lässt sich durch

$$\mathbb{V}(\hat{Y}) = \mathbb{E}(\hat{Y} - \mathbb{E}\hat{Y})^2 = \min_{\mu} \mathbb{E}(\hat{Y} - \mu)^2 \tag{3.1}$$

charakterisieren als ein Maß für die erwartete Distanz der Zufallsvariable $\hat{Y}$ zu seiner nächst entfernten deterministischen Zahl $\mathbb{E}\hat{Y}$, sowie das Bias

$$\mathrm{bias}(\hat{Y}) = \mathbb{E}\hat{Y} - \mathbb{E}Y, \ \mathrm{bias}^2(\hat{Y}) = (\mathbb{E}\hat{Y} - \mathbb{E}Y)^2 \tag{3.2}$$

als ein Maß für die Entfernung der beiden Erwartungswerte. Ist mit L die quadratische Verlustfunktion bezeichnet, so ergeben sich folgende Bestandteile des Testfehlers nach der üblichen Bias-Varianz-Zerlegung

$$\begin{aligned} \mathrm{Err}_z(\hat{Y}) &= \mathbb{E}L(\hat{Y}, Y) = \mathbb{E}(\hat{Y} - Y)^2 \\ &= \underbrace{\mathbb{V}(Y)}_{\text{irreduzibler Fehler}} + \underbrace{\mathrm{bias}^2(\hat{Y}) + \mathbb{V}(\hat{Y})}_{\text{reduzibler Fehler}}. \end{aligned} \tag{3.3}$$

Standardmäßige Umformungen des Varianzterms

$$\mathbb{V}(\hat{Y}) = \mathbb{E}((Y - \hat{Y})^2 - (Y - \mathbb{E}\hat{Y})^2) \tag{3.4}$$

und des Biasterms

$$\text{bias}^2(\hat{Y}) = \mathbb{E}((Y - \mathbb{E}\hat{Y})^2 - (Y - \mathbb{E}Y)^2) \tag{3.5}$$

ergeben eine alternative Darstellung der Bias-Varianz-Zerlegung

$$\text{Err}_z(\hat{Y}) = \underbrace{\mathbb{E}L(Y, \mathbb{E}Y)}_{\mathbb{V}(Y)} + \underbrace{\mathbb{E}(L(Y, \mathbb{E}\hat{Y}) - L(Y, \mathbb{E}Y))}_{\text{bias}^2(\hat{Y})} + \underbrace{\mathbb{E}(L(Y, \hat{Y}) - L(Y, \mathbb{E}\hat{Y}))}_{\mathbb{V}(\hat{Y})}. \tag{3.6}$$

Durch die Wahl des Regressionsmodells kann nur Einfluss auf den reduziblen Fehler genommen werden. Die Varianz dient zum einen dem Zweck, ein Maß für die Variabilität von $\hat{Y}$ zu sein, und zum anderen den Testfehler wie in (3.3) zu erklären.
Motiviert durch (3.1) und (3.2), verallgemeinert James [8, S. 120] die Definitionen für beliebige symmetrische Verlustfunktionen.

Definition 3.1. Sei L eine symmetrische Verlustfunktion auf $\mathcal{G} \times \mathcal{G}$, i. e. $L(k, k') = L(k', k)$ für alle $k, k' \in \mathcal{G}$ und $\hat{G} = \hat{G}(X)$ eine an den Datensatz z angepasste Klassifikationsregel, so ist der *systematische Bestandteil* von $\hat{G}$ (*systematic part*) definiert als

$$\mathbb{S}\hat{G} = \arg\min_{\mu} \mathbb{E}L(\hat{G}, \mu),$$

sowie analog

$$\mathbb{S}G = \arg\min_{\mu} \mathbb{E}L(G, \mu).$$

Die *Varianz eines Klassifikators* bzgl. L ist definiert durch

$$\mathbb{V}(\hat{G}) = \mathbb{E}L(\hat{G}, \mathbb{S}\hat{G}),$$

die *Varianz einer kategorialen Zufallsvariable* analog durch

$$\mathbb{V}(G) = \mathbb{E}L(G, \mathbb{S}G),$$

und das *Bias* als

$$\text{bias}(\hat{G}) = L(\mathbb{S}G, \mathbb{S}\hat{G}).$$

Ein Klassifikator $\hat{G}$ heißt *erwartungstreu*, falls $\mathbb{S}G = \mathbb{S}\hat{G}$ gilt. Eine sinnvolle zusätzliche Eigenschaft einer Verlustfunktion ist so $L(k, k) = 0$ für alle $k \in \mathcal{G}$, damit im Falle der Erwartungstreue das Bias null ist.

Durch die Varianz wird also die erwartete Abweichung bzgl. L von $\hat{G}$ zu seinem nächsten deterministischen Wert $\mathbb{S}\hat{G}$ gemessen. Im reellen Fall und bei einer quadratischen Verlustfunktion ist der systematische Bestandteil der Erwartungswert und bei der euklidischen Metrik als Verlustfunktion gerade der Median der Verteilung. Eine additive Zerlegung des Testfehlers mit diesen Begriffen ist nicht möglich, wie James [8, S. 123-124] durch ein Beispiel beweist. Wie in Gleichung (3.6) führt die Definition des *Varianz-Effekts*

$$VE(\hat{G}, G) = \mathbb{E}(L(G, \hat{G}) - L(G, \mathbb{S}\hat{G}))$$

und des *systematischen Effekts*

$$SE(\hat{G}, G) = \mathbb{E}(L(G, \mathbb{S}\hat{G}) - L(G, \mathbb{S}G))$$

zu einer additiven Zerlegung des Testfehlers

$$\begin{aligned} \mathrm{Err}_z(\hat{G}) &= \mathbb{E}L(G, \hat{G}) \\ &= \mathbb{E}L(G, \mathbb{S}G) + \mathbb{E}(L(G, \mathbb{S}\hat{G}) - L(G, \mathbb{S}G)) + \mathbb{E}(L(G, \hat{G}) - L(G, \mathbb{S}\hat{G})) \\ &= \underbrace{\mathbb{V}(G)}_{\text{irreduzibler Fehler}} + \underbrace{SE(\hat{G}, G) + VE(\hat{G}, G)}_{\text{reduzibler Fehler}}. \end{aligned} \tag{3.7}$$

Betrachtet man zusätzlich die Modelle als abhängig von dem zufälligen Trainingsdatensatz, so ergeben die analogen Definitionen von Varianz, Bias, Varianz-Effekt und systematischem Effekt eine Zerlegung des erwarteten Testfehlers. Für diesen Fall, also für $\hat{G} = \hat{G}(X, Z)$ hat James seinen Bias-Varianz-Formalismus entworfen.
Die Entscheidungsregel mit kleinstem Testfehler ist der *Bayes-Klassifikator*. Er ist für eine gegebene Verlustfunktion definiert durch

$$\begin{aligned} G^*(x) &= \arg\min_{k \in \mathcal{G}} \mathbb{E}\left[L(G, k) | X = x\right] \\ &= \arg\min_{k \in \mathcal{G}} \sum_{g \in \mathcal{G}} L(g, k) \mathbb{P}(G = g | X = x) = \mathbb{S}(G | X = x). \end{aligned} \tag{3.8}$$

Wählt man die 0/1-Verlustfunktion, so ergibt sich mit

$$\begin{aligned} G^*(x) &= \arg\min_{k \in \mathcal{G}} \mathbb{P}(G \neq k | X = x) \\ &= \arg\max_{k \in \mathcal{G}} \mathbb{P}(G = k | X = x) = \arg\max_{k \in \mathcal{G}} p_k(x) \end{aligned} \tag{3.9}$$

gerade die Bayes-Entscheidungsregel wie sie in (2.1) definiert ist. Für einen weiteren beliebigen Klassifikator $\hat{G}(x)$ gilt somit

$$\mathbb{E}[L(G, G^*(x)) | X = x] \leq \mathbb{E}[L(G, \hat{G}(x) | X = x] = \sum_{g \in \mathcal{G}} L(g, \hat{G}(x)) \mathbb{P}(G = g | X = x) \tag{3.10}$$

und damit durch iterierte Erwartungswertbildung die Optimalitätseigenschaft

$$\mathrm{Err}_z(G^*) = \mathrm{Err}(G^*) \leq \mathrm{Err}_z(\hat{G}). \tag{3.11}$$

Der Bayes-Klassifikator ist dabei unabhängig von einem Trainingsdatensatz und ist nur theoretisch unter der Kenntnis der A-posteriori-Wahrscheinlichkeiten bekannt. Im Folgenden wird daher auch die Notation des erwarteten Testfehlers anstatt des Testfehlers für Entscheidungsregeln verwendet.
Domingos [6] definiert ähnliche Begriffe von unterschiedlicher Bedeutung. Der *primäre Klassifikator* (*main prediction*)

$$\hat{G}^*(x) = \arg\min_{k \in \mathcal{G}} \mathbb{E}L(\hat{G}(x, Z), k) = \mathbb{S}(\hat{G}(x, Z)),$$

lässt sich als die zu erwartende Entscheidungsregel der Klassifikationsmethode interpretieren. Er ist der Klassifikator, dessen erwarteter Verlust relativ zur Methode $\hat{G}(x, Z)$ minimal ist. Genauso wie bei der Bayes-Klassifikationsregel ist der Testfehler des primären Klassifikators von keiner Realisierung des Trainingsdatensatzes abhängig, wobei zugleich der Klassifikator selbst sehr wohl von der Verteilung des Trainingsdatensatzes abhängt. Domingos Varianz-Begriff beschreibt die Variabilität der Methode um diese Entscheidungsregel.

Definition 3.2. Das *Bias einer Stichprobe* x nach Domingos ist definiert durch

$$B(x) = L(G^*(x), \hat{G}^*(x)),$$

die *Varianz* durch

$$V(x) = \mathbb{E}L(\hat{G}^*(x), \hat{G}(x, Z)).$$

Der Erwartungswert $\mathbb{E}B(X)$ heißt *erwarteter Bias* und $\mathbb{E}V(X)$ *erwartete Varianz*. Das *Rauschen einer Stichprobe* x

$$N(x) = \mathbb{E}\left[L(G, G^*(x)|X = x)\right]$$

ist der unvermeidbare Verlust, der dadurch entsteht, dass die Zufallsvariable $G|X = x$ durch die deterministische und im obigen Sinne optimale Funktion $G^*(x)$ prognostiziert wird.

Der Erwartungswert des Rauschens ist gerade der Testfehler des Bayes-Klassifikators, die so genannte *Bayes-Fehlerrate*

$$\mathbb{E}N(X) = \mathbb{E}L(G, G^*(X)) = \mathrm{Err}(G^*).$$

Die erwartete Varianz ist hier kein Maß für den erwarteten Abstand zwischen einer Zufallsvariable und einem deterministischen Wert, sondern vielmehr der zwischen zwei Zufallsgrößen. James kommt daher zu dem Schluss, dass seine Begriffe vorzuziehen sind, da sie insbesondere Verallgemeinerungen der gewohnten Begriffe im reellen Fall sind. Dennoch definiert Domingos hier durch den primären Klassifikator einen brauchbaren Begriff. Um beide Theorien miteinander zu vereinen, wird hier mit Hilfe eines *methodenbedingten Effekts*

$$ME(\hat{G}, G) = \mathbb{E}(L(G, \hat{G}(X, Z)) - L(G, \hat{G}^*))$$

folgende Zerlegung vorgeschlagen:

$$\begin{aligned}
\mathrm{Err}(\hat{G}) &= \mathbb{E}L(G, \hat{G}(X, Z)) \\
&= \mathbb{E}L(G, \mathbb{S}G) + \mathbb{E}(L(G, \mathbb{S}\hat{G}^*) - L(G, \mathbb{S}G)) \\
&\quad + \mathbb{E}(L(G, \hat{G}^*) - L(G, \mathbb{S}\hat{G}^*)) + \mathbb{E}(L(G, \hat{G}(X, Z)) - L(G, \hat{G}^*)) \qquad (3.12) \\
&= \mathbb{V}(G) + SE(\hat{G}^*, G) + VE(\hat{G}^*, G) + ME(\hat{G}, G) \\
&= \mathrm{Err}(\hat{G}^*) + ME(\hat{G}, G).
\end{aligned}$$

Diese Darstellung des erwarteten Testfehlers hat den Vorteil, dass sie die Variabilität der Methode um den primären Klassifikator abbildet.

Beispiel 3.3. Angelehnt an James [8, S. 124 und 125] dienen zur Illustration der Begriffe die Verteilungen und Klassifikatoren aus Tabelle 1. Dabei wird für dieses Beispiel die 0/1-Verlustfunktion verwendet. Der systematische Bestandteil ist in diesem Fall der Modalwert.

$$\mathbb{S}(\hat{G}) = \arg\min_{k\in\mathcal{G}} \mathbb{E}I(\hat{G} \neq k) = \arg\max_{k\in\mathcal{G}} \mathbb{P}(\hat{G} = k)$$

Varianz und Bias einer Methode, bzw. eines Klassifikators sind durch

$$\begin{aligned} \mathbb{V}(\hat{G}) &= \mathbb{P}(\hat{G} \neq \mathbb{S}\hat{G}) = 1 - \max_{k\in\mathcal{G}} \mathbb{P}(\hat{G} = k), \\ \text{bias}(\hat{G}) &= I(\mathbb{S}\hat{G} \neq \mathbb{S}G) \end{aligned} \tag{3.13}$$

gegeben, Varianz-Effekt und systematischer Effekt durch

$$\begin{aligned} VE(\hat{G}, G) &= \mathbb{P}(G \neq \hat{G}) - \mathbb{P}(G \neq \mathbb{S}\hat{G}), \\ SE(\hat{G}, G) &= \mathbb{P}(G \neq \mathbb{S}\hat{G}) - \mathbb{P}(G \neq \mathbb{S}G). \end{aligned} \tag{3.14}$$

k	**1**	**2**	**3**
$\mathbb{P}(G = k)$	0.5	0.4	0.1
x	**4**	**5**	**6**
$\mathbb{P}(X = x)$	0.1	0.5	0.4
$\mathbb{P}(G = 1, X = x)$	0.03	0.15	0.32
$\mathbb{P}(G = 2, X = x)$	0.06	0.30	0.04
$\mathbb{P}(G = 3, X = x)$	0.01	0.05	0.04
$\mathbb{P}(G = 1 \mid X = x)$	0.3	0.3	0.8
$\mathbb{P}(G = 2 \mid X = x)$	0.6	0.6	0.1
$\mathbb{P}(G = 3 \mid X = x)$	0.1	0.1	0.1
$\hat{G}_A(x)$	3	2	1
$\hat{G}_B(x)$	1	2	3
$\hat{G}_C(x)$	1	1	1
$G^*(x)$	2	2	1

Tabelle 1: Rechenbeispiel eines Drei-Klassen-Problems, $\mathcal{G} = \{1, 2, 3\}$, $\mathcal{X} = \{4, 5, 6\}$

Die berechneten Größen von Bias, Lage- und Streuungsmaße der kategorialen Zufallsvariablen aus Tabelle 1

$$\begin{aligned} &\mathbb{S}G = \mathbb{S}\hat{G}_C = 1,\ \mathbb{S}G^* = \mathbb{S}\hat{G}_A = \mathbb{S}\hat{G}_B = 2, \\ &\mathbb{V}(G) = \mathbb{V}(\hat{G}_A) = \mathbb{V}(\hat{G}_B) = 0.5,\ \mathbb{V}(G^*) = 0.4,\ \mathbb{V}(\hat{G}_C) = 0, \\ &\text{bias}(\hat{G}_A) = \text{bias}(\hat{G}_B) = \text{bias}(G^*) = 1,\ \text{bias}(\hat{G}_C) = 0, \end{aligned}$$

sowie die Varianz-Effekte und die systematischen Effekte

$$
\begin{aligned}
&VE(\hat{G}_A, G) = 0.09 + 0.2 + 0.08 - 0.6 = -0.23,\\
&VE(\hat{G}_B, G) = 0.07 + 0.2 + 0.36 - 0.6 = 0.03,\\
&VE(\hat{G}_C, G) = 0,\\
&VE(G^*, G) = 0.04 + 0.2 + 0.08 = -0.28,\\
&SE(\hat{G}_A, G) = SE(\hat{G}_B, G) = SE(G^*, G) = \mathbb{P}(G \neq 2) - \mathbb{P}(G \neq 1) = 0.1,\\
&SE(\hat{G}_C, G) = 0,
\end{aligned}
$$

erklären nach (3.7) die Ursachen des Testfehlers:

$$
\begin{aligned}
&\mathrm{Err}(\hat{G}_A) = 0.37 = 0.5 + 0.1 - 0.23,\\
&\mathrm{Err}(\hat{G}_B) = 0.63 = 0.5 + 0.1 + 0.03,\\
&\mathrm{Err}(\hat{G}_C) = 0.5,\\
&\mathrm{Err}(G^*) = 0.32 = 0.5 + 0.1 - 0.28.
\end{aligned}
$$

Der Klassifikator $\hat{G}_C$ ordnet jede Beobachtung der ersten Gruppe hinzu, dem Modalwert der Zufallsvariable G. Aufgrund dieser deterministischen Vorgehensweise ist sowohl die Varianz als auch der Varianz-Effekt gleich null. Indem man es einem Klassifikator erlaubt, von der Zufallsvariablen X abzuhängen, erhöht sich die Varianz und dennoch kann sich der Fehler im Falle eines negativen Varianz-Effekts verringern. Eine Aussage, die nicht für den reellen Standardfall gilt, da hier Varianz und Varianz-Effekt des Regressionsmodells gleich sind. Die Entscheidungsregel $\hat{G}_B$ trifft jedoch für $X = 6$ die falsche Wahl und erhöht dadurch seinen Testfehler. Varianz und Bias sind per Definition nichtnegativ. Zudem folgt aus der Definition des systematischen Bestandteils durch

$$\mathbb{E}L(G, \mathbb{S}\hat{G}) \geq \mathbb{E}L(G, \mathbb{S}G),$$

dass auch der systematische Effekt nichtnegativ ist. Konstruiert man nun eine Methode, die an einen Datensatz, der aus nur einer Beobachtung besteht, angepasst wird, so kann Z neun verschiedene Werte annehmen. Definiert man zum Beispiel die Methoden

$$
\hat{G}_D(x, z) = \begin{cases} \hat{G}_A(x), \text{ für } z \in A \\ G^*(x), \text{ für } z \in B \end{cases} \qquad \hat{G}_E(x, z) = \begin{cases} \hat{G}_A(x), \text{ für } z \in A \\ \hat{G}_B(x), \text{ für } z \in B \end{cases}
$$

$$
\text{mit } A = \{(4,3),(5,2),(6,1)\},\ B = \{(4,1),(4,2),(5,1),(5,3),(6,2),(6,3)\},
$$

$$
\text{und } \mathbb{P}(Z \in A) = 0.63,\ \mathbb{P}(Z \in B) = 0.37,
$$

so ergeben sich unterschiedliche Verteilungen (s. Tabelle 2). Dabei gehen die Methoden so vor, dass sie $\hat{G}_A$ wählen, falls sie den beobachteten Datensatz richtig klassifiziert. In allen anderen Fällen gilt die jeweils andere Entscheidungsregel. Die Wahrscheinlichkeit von $\{Z \in B\}$ ist demnach gerade der Testfehler von $\hat{G}_A$.

k	**1**	**2**	**3**
$\mathbb{P}(\hat{G}_D(X,Z)=k)$	0.4	0.537	0.063
$\mathbb{P}(\hat{G}_E(X,Z)=k)$	0.289	0.5	0.211

Tabelle 2: Verteilung der Klassifikationsmethoden

Eine Analyse der Variabilität der Methoden um deterministische Werte durch James' Kenngrößen liefert:

$$\begin{aligned}
&\mathbb{S}\hat{G}_D = \mathbb{S}\hat{G}_E = 2,\ \text{bias}(\hat{G}_D) = \text{bias}(\hat{G}_E) = 1,\ SE(\hat{G}_D) = SE(\hat{G}_E) = 0.1,\\
&\mathbb{V}(\hat{G}_D(X,Z)) = 0.463,\ \mathbb{V}(\hat{G}_E(X,Z)) = 0.5,\\
&VE(\hat{G}_D, G) = 0.63 \cdot VE(\hat{G}_A, G) + 0.37 \cdot VE(G^*, G) = -0.2484,\\
&VE(\hat{G}_E, G) = 0.63 \cdot VE(\hat{G}_A, G) + 0.37 \cdot VE(\hat{G}_B, G) = -0.1338,\\
&\text{Err}(\hat{G}_D) = 0.3516 = 0.5 + 0.1 - 0.2484,\\
&\text{Err}(\hat{G}_E) = 0.4662 = 0.5 + 0.1 - 0.1338.
\end{aligned}$$

Für den Varianz-Effekt gilt dabei

$$\begin{aligned}
VE(\hat{G}_D, G) &= \mathbb{P}(G \neq \hat{G}_D(X,Z)) - \mathbb{P}(G \neq \mathbb{S}\hat{G}_D)\\
&= \mathbb{P}(G \neq \hat{G}_A(X), Z \in A) + \mathbb{P}(G \neq G^*(X), Z \in B) - \mathbb{P}(G \neq 2) \qquad (3.15)\\
&= \mathbb{P}(Z \in A) \cdot VE(\hat{G}_A, G) + \mathbb{P}(Z \in B) \cdot VE(G^*, G).
\end{aligned}$$

Aus

$$\mathbb{P}(\hat{G}_D(x,Z)=k) = \mathbb{P}(Z \in A) \cdot I(\hat{G}_A(x)=k) + \mathbb{P}(Z \in B) \cdot I(G^*(x)=k)$$

und der Tatsache, dass das Ereignis $\{Z \in A\}$ wahrscheinlicher ist als sein komplementäres, folgt für den primären Klassifikator

$$\hat{G}_D^*(x) = \underset{k \in \mathcal{G}}{\arg\min}\, \mathbb{P}(\hat{G}_D(x,Z) \neq k) = \underset{k \in \mathcal{G}}{\arg\max}\, \mathbb{P}(\hat{G}_D(x,Z)=k) = \hat{G}_A(x). \qquad (3.16)$$

x	**4**	**5**	**6**
$B_D(x) = B_E(x)$	1	0	0
$V_D(x)$	0.37	0	0
$V_E(x)$	0.37	0	0.37
$N(x)$	0.4	0.4	0.2

Tabelle 3: Bias, Varianz und Rauschen nach Domingos

Die Untersuchung der Variabilität der Methode um den primären Klassifikator führt

daher zu folgenden Ergebnissen:

$$\begin{aligned}
&\mathbb{E}B_D(X) = \mathbb{E}B_E(X) = 0.1,\ \mathbb{E}V_D(X) = 0.037,\ \mathbb{E}V_E(X) = 0.185,\ \mathbb{E}N(X) = 0.32,\\
&ME(\hat{G}_D, G) = 0.63 \cdot (\mathrm{Err}(\hat{G}_A) - \mathrm{Err}(\hat{G}_A)) + 0.37 \cdot (\mathrm{Err}(G^*) - \mathrm{Err}(\hat{G}_A)) = -0.0185,\\
&ME(\hat{G}_E, G) = 0.37(\mathrm{Err}(\hat{G}_B) - \mathrm{Err}(\hat{G}_A)) = 0.0962,\\
&\mathrm{Err}(\hat{G}_D) = 0.3515 = \mathrm{Err}(\hat{G}_D^*) + ME(\hat{G}_D, G) = 0.37 - 0.0185,\\
&\mathrm{Err}(\hat{G}_E) = 0.4662 = 0.37 + 0.0962.
\end{aligned}$$

Die Errechnung des methodenbedingten Effekts erfolgt gemäß

$$\begin{aligned}
ME(\hat{G}_D, G) &= \mathbb{E}\mathbb{E}[L(G, \hat{G}_D(X, Z)) - L(G, \hat{G}_D^*)|Z = z]\\
&= \mathbb{P}(Z \in A) \cdot \mathbb{E}[L(G, \hat{G}_D(X, Z)) - L(G, \hat{G}_D^*)|Z \in A]\\
&\quad + \mathbb{P}(Z \in B) \cdot \mathbb{E}[L(G, \hat{G}_D(X, Z)) - L(G, \hat{G}_D^*)|Z \in B] \qquad (3.17)\\
&= \mathbb{P}(Z \in A) \cdot (\mathrm{Err}(\hat{G}_A) - \mathrm{Err}(\hat{G}_A))\\
&\quad + \mathbb{P}(Z \in B) \cdot (\mathrm{Err}(G^*) - \mathrm{Err}(\hat{G}_A)).
\end{aligned}$$

Für die Methode $\hat{G}_E$ gelten zu (3.15), (3.16) und (3.17) analoge Aussagen. Die erwartete Varianz nach Domingos von $\hat{G}_E$ ist größer als die von $\hat{G}_D$, da sich der Klassifikator $\hat{G}_B$ in gleich zwei Argumenten von $\hat{G}_A$ unterscheidet und er daher vom primären Klassifikator stärker abweicht. Durch den methodenbedingten Effekt von $\hat{G}_D$ verbessert sich die Fehlerrate, da sich die Methode in den Fällen $z \in B$ für den Bayes-Klassifikator entscheidet.

Domingos schlägt eine Zerlegung des Fehlers mit Hilfe seiner Definitionen vor. Für den Spezialfall einer 0/1-Verlustfunktion benutzt er Koeffizienten für die Zerlegung, von denen jedoch James [8, S. 128] zeigt, dass diese selbst Funktionen von Bias und Varianz sind. Zudem stellt Domingos die Aufgabe, seine Aussagen für beliebige Verlustfunktionen zu verallgemeinern. Von Satz 3.6 wird gezeigt, dass dieser eine Verallgemeinerung von Domingos Zerlegung darstellt. Er verdeutlicht noch einmal, dass diese nicht sinnvoll ist, da seine Koeffizienten Funktionen von Bias und Varianz sind.

Satz 3.4 (Domingos)**.** Sei L die 0/1-Verlustfunktion, so gilt für das Klassifikationsproblem mit $K \geq 2$ und für eine beobachtete Stichprobe x die Zerlegung

$$\mathbb{E}[L(G, \hat{G}(x, Z))|X = x] = c_1(x)N(x) + B(x) + c_2(x)V(x)$$

mit $c_1(x) = \mathbb{P}(\hat{G}(x, Z) = G^*(x)) - \mathbb{P}(\hat{G}(x, Z) \neq G^*(x))\mathbb{P}(\hat{G}(x, Z) = G|G^*(x) \neq G)$ und $c_2(x) = 1$ falls $\hat{G}^*(x) = G^*(x)$, bzw. $c_2(x) = -\mathbb{P}(\hat{G}(x, Z) = G^*(x)|\hat{G}(x, Z) \neq \hat{G}^*(x))$ falls $\hat{G}^*(x) \neq G^*(x)$ gilt.

Beweis. Siehe Domingos [6]. □

Korollar 3.5. Für den erwarteten Testfehler einer Klassifikationsregel gilt dann

$$\mathrm{Err}(\hat{G}) = \mathbb{E}[\mathbb{E}[L(G, \hat{G}(x, Z))|X = x]] = \mathbb{E}[c_1(X)N(X)] + \mathbb{E}[B(X)] + \mathbb{E}[c_2(X)V(X)].$$

Sei $C = (c_{ij})_{i,j\in\mathcal{G}}$ die zu L gehörige Verlustmatrix mit $c_{ij} = L(i,j)$ für $i,j \in \mathcal{G}$ und mit e_i der i-te Standardvektor bezeichnet, so gilt $e_i^T C e_j = L(i,j)$ für alle $i,j \in \mathcal{G}$. Mit den Vektoren

$$\begin{aligned} p(x) &= \mathbb{E}[e_G|X=x] = (p_1(x),\dots,p_K(x))^T, \\ q(x) &= \mathbb{E}[e_{\hat{G}(x,Z)}] = (\mathbb{P}(\hat{G}(x,Z)=1),\dots,\mathbb{P}(\hat{G}(x,Z)=K))^T, \end{aligned} \tag{3.18}$$

kann so der Fehler im Sinne Domingos für eine beliebige Verlustfunktion zerlegt werden.

Satz 3.6. Angenommen für eine Klassifikationsmethode $\hat{G}(x,Z)$, für eine beliebige Verlustfunktion L und für einen Eingabewert $x \in \mathcal{X}$ gilt $V(x) \neq 0$ und $N(x) \neq 0$, so folgt aus der Definition der Koeffizienten

$$c_0(x,z) = \frac{p(x)^T C e_{\hat{G}(x,z)} - e_{G^*(x)}^T C e_{\hat{G}(x,z)}}{p(x)^T C e_{\hat{G}(x)}} = \frac{p(x)^T C e_{\hat{G}(x,z)} - e_{G^*(x)}^T C e_{\hat{G}(x,z)}}{N(x)},$$

$$c_1(x) = \mathbb{E}[c_0(x,Z)] \text{ und } c_2(x) = \frac{e_{G^*(x)}^T C q(x) - e_{G^*(x)}^T C e_{\hat{G}^*(x)}}{e_{\hat{G}^*(x)}^T C q(x)} = \frac{e_{G^*(x)}^T C q(x) - B(x)}{V(x)},$$

die Zerlegung

$$\mathbb{E}[L(G,\hat{G}(x,Z))|X=x] = c_1(x)N(x) + B(x) + c_2(x)V(x).$$

Beweis.
1. Schritt:

$$\begin{aligned} \mathbb{E}[L(G,\hat{G}(x,z))|X=x] &= \mathbb{E}[e_G^T C e_{\hat{G}(x,z)}|X=x] = p(x)^T C e_{\hat{G}(x,z)} \\ &= c_0(x,z) p(x)^T C e_{G^*(x)} + e_{G^*(x)^T} C e_{\hat{G}(x,z)} \\ &= c_0(x,z)\mathbb{E}[L(G,G^*(x))|X=x] + L(G^*(x),\hat{G}(x,z)) \\ &= c_0(x,z)N(x) + L(G^*(x),\hat{G}(x,z)) \end{aligned} \tag{3.19}$$

2. Schritt:

$$\begin{aligned} \mathbb{E}[L(G^*(x),\hat{G}(x,Z))] &= e_{G^*(x)}^T C \mathbb{E}[e_{\hat{G}(x,Z)}] = e_{G^*(x)}^T C q(x) \\ &= e_{G^*(x)}^T C e_{\hat{G}^*(x)} + c_2(x) e_{\hat{G}^*(x)} C q(x) \\ &= L(G^*(x),\hat{G}^*(x)) + c_2(x) e_{\hat{G}^*}^T C q(x) \\ &= B(x) + c_2(x)\mathbb{E}[L(\hat{G}^*(x),\hat{G}(x,Z))] \\ &= B(x) + c_2(x)V(x) \end{aligned} \tag{3.20}$$

3. Schritt, analog zu Domingos:

$$\begin{aligned} \mathbb{E}[L(G,\hat{G}(x,Z))|X=x] &= \mathbb{E}[\mathbb{E}[L(G,\hat{G}(x,z))|Z=z]|X=x] \\ &= \mathbb{E}[\mathbb{E}[c_0(x,z)N(x) + L(G^*(x),\hat{G}(x,z))|Z=z]|X=x] \\ &= \mathbb{E}[c_0(x,Z)]N(x) + \mathbb{E}[L(G^*(x),\hat{G}(x,Z))] \\ &= c_1(x)N(x) + B(x) + c_2(x)V(x) \end{aligned} \tag{3.21}$$

□

Bemerkung 3.7. Es bleibt noch zu zeigen, dass Satz 3.6 eine Verallgemeinerung des Satzes 3.4 ist. Im Folgenden sind die Wahrscheinlichkeiten bedingt nach $X = x$ zu verstehen, sofern die entsprechenden Zufallsvariablen überhaupt von X abhängig sind. Außerdem ist zu beachten, dass G und Z unabhängig sind. Es gilt

$$\begin{aligned}
\mathbb{P}(G = \hat{G}(Z)) &= \mathbb{P}(\hat{G}(Z) \neq G^*, G^* \neq G)\mathbb{P}(\hat{G}(Z) = G | G^* \neq G) \\
&\quad + \mathbb{P}(G^* = G)\mathbb{P}(G^* = \hat{G}(Z) | G^* = G) \\
&= \mathbb{P}(\hat{G}(Z) \neq G^*)\mathbb{P}(G^* \neq G)\mathbb{P}(\hat{G}(Z) = G | G^* \neq G) \\
&\quad + \mathbb{P}(G^* = \hat{G}(Z), G^* = G) \\
&= \mathbb{P}(\hat{G}(Z) \neq G^*)\mathbb{P}(\hat{G}(Z) = G, G^* \neq G) \\
&\quad + \mathbb{P}(G^* = \hat{G}(Z))\mathbb{P}(G^* = G).
\end{aligned} \tag{3.22}$$

Mit C sei die Verlustmatrix der $0/1$-Verlustfunktion bezeichnet, so gilt für den Koeffizent $c_1(x)$ von $N(x)$ nach Domingos

$$\begin{aligned}
c_1(x) &= \mathbb{P}(\hat{G}(Z) = G^*) - \mathbb{P}(\hat{G}(Z) \neq G^*)\mathbb{P}(\hat{G}(Z) = G | G^* \neq G) \\
&= \frac{\mathbb{P}(G^* = \hat{G}(Z))\mathbb{P}(G^* \neq G) - \mathbb{P}(\hat{G}(Z) \neq G^*)\mathbb{P}(\hat{G}(Z) = G, G^* \neq G)}{\mathbb{P}(G^* \neq G)} \\
&= \frac{\mathbb{P}(G^* = \hat{G}(Z))}{\mathbb{P}(G^* \neq G)} \\
&\quad - \frac{\mathbb{P}(\hat{G}(Z) \neq G^*)\mathbb{P}(\hat{G}(Z) = G, G^* \neq G) + \mathbb{P}(G^* = \hat{G}(Z))\mathbb{P}(G^* = G)}{\mathbb{P}(G^* \neq G)} \\
&= \frac{\mathbb{P}(G^* = \hat{G}(Z)) - \mathbb{P}(G = \hat{G}(Z))}{\mathbb{P}(G^* \neq G)} \\
&= \frac{-\mathbb{P}(G^* \neq \hat{G}(Z)) + \mathbb{P}(G \neq \hat{G}(Z))}{\mathbb{P}(G^* \neq G)} \\
&= \frac{p^T C q - e_{G^*}^T C q}{p^T C e_{G^*}}.
\end{aligned} \tag{3.23}$$

Für Domingos Koeffizient $c_2(x)$ von $V(x)$ gilt zudem

$$\begin{aligned}
\text{falls } \hat{G}^* = G^* : \frac{e_{G^*}^T C q - e_{G^*}^T C e_{G^*}}{e_{\hat{G}^*}^T C q} &= \frac{e_{\hat{G}^*}^T C q}{e_{\hat{G}^*}^T C q} = 1 = c_2(x), \\
\text{falls } \hat{G}^* \neq G^* : \frac{e_{G^*}^T C q - e_{G^*}^T C e_{\hat{G}^*}}{e_{\hat{G}^*}^T C q} &= \frac{\mathbb{P}(G^* \neq \hat{G}(Z)) - I(G^* \neq \hat{G}^*)}{\mathbb{P}(\hat{G}^* \neq \hat{G}(Z))} \\
&= \frac{1 - \mathbb{P}(G^* = \hat{G}(Z)) - 1}{\mathbb{P}(\hat{G}^* \neq \hat{G}(Z))} = \frac{-\mathbb{P}(G^* = \hat{G}(Z))}{\mathbb{P}(\hat{G}^* \neq \hat{G}(Z))} \\
&= \frac{-\mathbb{P}(G^* = \hat{G}(Z), \hat{G}(Z) \neq \hat{G}^*)}{\mathbb{P}(\hat{G}^* \neq \hat{G}(Z))} = -\mathbb{P}(\hat{G}(Z) = G^* | \hat{G}(Z) \neq \hat{G}^*) = c_2(x).
\end{aligned}$$

3.3 Kreuzvalidierung

Die Kreuzvalidierung ist ein Verfahren zur Schätzung des (erwarteten) Testfehlers. Der Datensatz wird dafür in n möglichst gleichgroße Partitionen geteilt. Für i=1,...,n wird jeweils ein Modell $\hat{G}^{-i}$ an den Trainingsdatensatz ohne die i-te Teildatenmenge angepasst. Die ausgelassene i-te Datenmenge wird dann zur Validierung des geschätzten Modells $\hat{G}^{-i}$ verwendet. Sei $\kappa : \{1, \ldots, N\} \to \{1, \ldots, n\}$ die Funktion, die der i-ten Beobachtung den Index ihrer Partition zuweist, so ist die *n-fache Kreuzvalidierung* definiert als

$$CV(\hat{G}) = \frac{1}{N} \sum_{i=1}^{N} L(g_i, \hat{G}^{-\kappa(i)}(x_i)).$$

Gilt $n = N$, so besteht jede Partition aus genau einem Datenpunkt. In diesem Fall heißt $CV(\hat{G})$ *Leave-One-Out-Kreuzvalidierung*. Die Komplexität einiger Modelle $\hat{G}(x, \alpha)$ ist oft durch einen Tuning-Parameter α bestimmt. Mit Hilfe der Kreuzvalidierung

$$CV(\hat{G}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(g_i, \hat{G}^{-\kappa(i)}(x_i, \alpha))$$

schätzt man die Kurve des erwarteten Testfehlers in Abhängigkeit von α. Gewählt wird dann der Tuning-Parameter $\hat{\alpha}$, der die geschätzte Kurve und damit den geschätzten zu erwartenden Testfehler minimiert, um die Komplexität des Modells adäquat zu bestimmen (s. Hastie, Tibshirani und Friedman [7, S. 241-242]).

4 Überwachte Klassifikationsverfahren

4.1 Lineare Diskriminanzanalyse (LDA)

Der Satz von Bayes schafft eine Verbindung zwischen den A-priori- und den A-posteriori-Wahrscheinlichkeiten. Definiert man die bedingte Dichte von X unter $G = k$ als

$$f_k(x) = f_{X|G=k}(x),$$

so führt dieser Satz auf die Gleichung

$$p_k(x) = \mathbb{P}(G = k \mid X = x) = \frac{\mathbb{P}(G = k) f_{X|G=k}(x)}{\sum_{l=1}^{K} \mathbb{P}(G = l) f_{X|G=l}(x)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{4.1}$$

Eine Modellierung und Schätzung von $p_k(x)$ ist also durch eine Parametrisierung und Schätzung der Dichten f_k und einer Schätzung der Wahrscheinlichkeiten π_k zu erreichen. In einigen Anwendungen sind die Eingabewerte, gegebenenfalls erst nach einer geeigneten Transformation, normalverteilt. Die lineare Diskriminanzanalyse macht daher die Annahme, dass die bedingte Verteilung von X unter $G = k$ einer multivariaten Normalverteilung entspricht mit der Dichte

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)},\ \mu_k \in \mathbb{R}^p,\ \Sigma_k \in \mathbb{R}^{p \times p} \tag{4.2}$$

und nimmt des Weiteren zur Vereinfachung an, dass die Kovarianzmatrizen Σ_k von der Klasse unabhängig sind.

$$\Sigma_k = \Sigma\ \forall k \in \mathcal{G} \tag{4.3}$$

Die Parameter und Wahrscheinlichkeiten lassen sich dann mit Hilfe des Trainingsdatensatzes schätzen:

$$\begin{aligned} \hat{\pi}_k &= N_k/N, \text{ wobei } N_k \text{ Beobachtungen zur Klasse } k \text{ gehören,} \\ \hat{\mu}_k &= \sum_{g_i = k} x_i / N_k, \\ \hat{\Sigma} &= \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K). \end{aligned} \tag{4.4}$$

Gilt für eine neue Beobachtung x die Ungleichung $0 < \log(p_k(x)/p_l(x))$, so ist es wahrscheinlicher, dass die Beobachtung zur Gruppe k gehört als zur Gruppe l. Die Umformung dieser Bedingung durch

$$\begin{aligned} 0 < \log \frac{p_k(x)}{p_l(x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) \\ &= x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k - (x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l + \log \pi_l) \end{aligned} \tag{4.5}$$

führt zur linearen Gleichung

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k = x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l + \log\pi_l,$$

die eine Hyperebene im p-dimensionalen Raum der Eingabewerte beschreibt; die *Entscheidungsgrenze*. Definiert man die *lineare Diskriminantenfunktion* als

$$\delta_k(x) = \delta_k(x;\mu_1,\ldots,\mu_K,\Sigma) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k, \tag{4.6}$$

so ist die Klassifikationsregel der Methode gegeben durch

$$\hat{G}(x) = \arg\max_{k\in\mathcal{G}} \delta_k(x;\hat{\mu}_1,\ldots,\hat{\mu}_K,\hat{\Sigma}). \tag{4.7}$$

Die Beobachtung wird also in die Gruppe k, für die $\log p_k(x)$ maximal ist, klassifiziert. Verzichtet man auf die Annahme der Gleichheit der Kovarianzmatrizen in (4.3), so führen analoge Rechnungen zu der *quadratische Diskriminantenfunktion*

$$\begin{aligned}\delta_k^2(x) &= \delta_k^2(x;\hat{\mu}_1,\ldots,\hat{\mu}_K,\hat{\Sigma}_1,\ldots,\hat{\Sigma}_K)\\ &= -\frac{1}{2}\log|\hat{\Sigma}_k| - \frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) + \log\hat{\pi}_k\\ &\text{mit } \hat{\Sigma}_k = \sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N_k-1)\end{aligned} \tag{4.8}$$

und damit zu einer Klassifikationsregel gemäß (4.7). Die Entscheidungsgrenzen werden dabei durch die quadratischen Gleichungen $\delta_k^2(x) = \delta_l^2(x)$ beschrieben. Die beiden Methoden werden von Hastie, Tibshirani und Friedman [7, S. 106-110] erläutert.

4.2 Multinomiale logistische Regression

Die multinomiale logistische Regression basiert auf dem Wunsch die logarithmierten Quotenverhältnisse

$$\log\frac{p_k(x)}{p_K(x)} = \log\frac{\mathbb{P}(G=k \mid X=x)}{\mathbb{P}(G=K \mid X=x)},\ k=1,\ldots,K-1$$

durch Regression an eine lineare Funktion $\beta_k^T x$ mit $\beta_k \in \mathbb{R}^{p+1}$ und $x=(1,x_1,\ldots,x_p)$ anzupassen. Die Idee ist also $K-1$ multiple lineare Regressionen durchzuführen, deren Modelle folgendermaßen gegeben sind

$$\begin{aligned}\log\frac{p_1(x)}{p_K(x)} &= \beta_1^T x,\\ &\vdots\\ \log\frac{p_{K-1}(x)}{p_K(x)} &= \beta_{K-1}^T x,\\ \text{mit } \sum_{k=1}^K p_k(x) &= 1.\end{aligned} \tag{4.9}$$

Dieses Vorgehen ist jedoch nicht umsetzbar, da die logarithmierten Quotenverhältnisse nicht beobachtet werden können. Es gilt ja gerade die A-posteriori-Wahrscheinlichkeiten zu schätzen. Elementare Umformungen von (4.9) führen daher auf das multinomiale logistische Regressionsmodell

$$\begin{aligned} p_1(x) &= \frac{\exp\left(\beta_1^T x\right)}{1+\sum_{l=1}^{K-1}\exp\left(\beta_l^T x\right)} \\ &\vdots \\ p_{K-1}(x) &= \frac{\exp\left(\beta_{K-1}^T x\right)}{1+\sum_{l=1}^{K-1}\exp\left(\beta_l^T x\right)} \\ p_K(x) &= \frac{1}{1+\sum_{l=1}^{K-1}\exp\left(\beta_l^T x\right)}. \end{aligned} \tag{4.10}$$

Die Wahrscheinlichkeiten $p_k\left(x;\vartheta\right) = \mathbb{P}_\vartheta\left(G=k \mid X=x\right)$ werden so durch

$$\vartheta = \left(\beta_1^T, \ldots, \beta_{K-1}^T\right) \in \mathbb{R}^{(1+p)(K-1)}$$

parametrisiert. Liegt ein Trainingsdatensatz (x_i, g_i) von N Beobachtungen vor, so lässt sich ein Maximum-Likelihood-Schätzer $\hat{\vartheta}$ durch das Lösen der Score-Gleichung

$$\frac{\partial l\left(\vartheta\right)}{\partial \vartheta} = 0$$

finden, während die bedingte Log-Likelihood-Funktion definiert ist durch

$$l\left(\vartheta\right) = \sum_{i=1}^{N} \log p_{g_i}\left(x_i;\vartheta\right).$$

Die Schätzung ist dabei unabhängig von der gewählten K-ten Vergleichsklasse. Mit Hilfe des Maximum-Likelihood-Schätzers lässt sich dann eine Klassifikationsregel wie in (2.2) formulieren. Diese Vorgehensweise wird so von Hastie, Tibshirani und Friedman [7, S. 119-120] erläutert.

Bemerkung 4.1. In Kapitel 4.1 folgte bereits aus den Annahmen der linearen Diskriminanzanalyse (4.2) und (4.3), dass die logarithmierten Quotenverhältnisse linear in x sind.

$$\begin{aligned} \delta_k(x) - \delta_K(x) &= \log\frac{p_k(x)}{p_K(x)} \\ &= \log\frac{\pi_k}{\pi_K} - \frac{1}{2}\left(\mu_k-\mu_K\right)^T\Sigma^{-1}\left(\mu_k+\mu_K\right) + \left(\mu_k-\mu_K\right)^T\Sigma^{-1}x \end{aligned} \tag{4.11}$$

Die logistische Regression setzt schwächere Bedingungen als die LDA voraus. Insbesondere wird auf eine Parametrisierung der Verteilung von X unter der Bedingung $G=k$ verzichtet. Dennoch liefert die lineare Diskriminanzanalyse in der Praxis laut Hastie, Tibshirani und Friedman [7, S. 127-128] vergleichbar gute Ergebnisse, selbst wenn $X|G=k$ die Normalverteilungsannahme nicht erfüllt und sogar wenn die Eingabewerte nicht einmal stetig verteilt sind.

4.3 Support-Vector-Machine (SVM)

4.3.1 Optimal trennende Hyperebene und C-Support-Vector-Klassifikator (C-SVC)

Zur Klassifizierung der Daten in zwei Gruppen trennen sowohl die LDA als auch die logistische Regression den Eingaberaum linear, d. h. durch eine Hyperebene

$$H = \left\{x \in \mathcal{X} = \mathbb{R}^p : \ x^T\beta + \beta_0 = 0\right\}, \ \beta_0 \in \mathbb{R}, \ \beta \in \mathbb{R}^p \text{ o. B. d. A. } ||\beta|| = 1$$

in zwei Hälften

$$\mathcal{X}_1 = \{x \in \mathbb{R}^p : \ x^T\beta + \beta_0 > 0\} \text{ und } \mathcal{X}_2 = \{x \in \mathbb{R}^p : \ x^T\beta + \beta_0 < 0\}.$$

Dabei werden β und β_0 bestimmt, indem die jeweiligen Parameter der Verteilung von $G|X = x$ geschätzt werden. Die Grundidee einer *Support-Vector-Machine* (SVM) ist es, die Hyperebene H geometrisch durch das Lösen eines Optimierungsproblems zu bestimmen, ohne dabei eine Verteilungsannahme über die A-posteriori-Wahrscheinlichkeiten zu treffen. Zudem können durch den Kernel-Trick die Eingabewerte in einen höher dimensionalen Raum transformiert werden, in dem man sich eine bessere lineare Trennbarkeit erhofft. Die rücktransformierte Entscheidungsgrenze ist dabei im Allgemeinen nichtlinear (s. Kapitel 4.3.3). Gilt $K = 2$, so lässt sich der Ausgabewert $g_1 = 1$ mit $y_1 = 1$ und $g_2 = 2$ mit $y_2 = -1$ codieren. Eine trennende Hyperebene, deren induzierte Klassifikationsregel

$$G(x) = \text{sign}\left(x^T\beta + \beta_0\right) \tag{4.12}$$

jedem Eingabewert x_i seinen tatsächlichen Ausgabewert y_i zuordnet, kann nur gefunden werden, falls ein $\beta \in \mathbb{R}^p$ und ein $\beta_0 \in \mathbb{R}$ existiert mit

$$y_i(x_i^T\beta + \beta_0) > 0 \text{ für alle } i = 1, \ldots, N. \tag{4.13}$$

Die geometrische Motivation der SVM ist es, den kleinsten aller Abstände R zwischen den Datenpunkten und der Hyperebene zu maximieren, damit der Bereich um die Hyperebene

$$B = \{x \in \mathbb{R}^p : \inf_{h \in H} ||x - h|| < R\},$$

in dem keine Eingabewerte liegen, möglichst groß ist. Man spricht dann von einer *optimal trennenden Hyperebene*. Das entsprechende Optimierungsproblem lautet

$$\begin{aligned} &\max_{\beta_0, \beta, ||\beta||=1} R \\ &\text{unter der Nebenbedingung } y_i(x_i^T\beta + \beta_0) \geq R \ \text{ für alle } i = 1, \ldots, N. \end{aligned} \tag{4.14}$$

Verzichtet man auf die Normierungsbedingung $||\beta|| = 1$, so ergibt sich die Nebenbedingung in (4.14) zu $y_i(x_i^T\beta/||\beta|| + \beta_0) \geq R$. Durch die Substitution $R = 1/||\beta||$ ist das Optimierungsproblem äquivalent zu

$$\begin{aligned} &\min_{\beta_0, \beta} ||\beta|| \\ &\text{unter der Nebenbedingung } y_i(x_i^T\beta + \beta_0) \geq 1 \ \text{ für alle } i = 1, \ldots, N. \end{aligned} \tag{4.15}$$

Sind die Daten nicht trennbar, gilt also (4.13) nicht, so können nichtnegative Schlupfvariablen $\xi = (\xi_1, \ldots, \xi_N)$ eingeführt werden, sodass $y_i(x_i^T\beta + \beta_0) \geq R(1 - \xi_i)$ für alle i dennoch erfüllt ist. Dem i-ten Datenpunkt wird es so erlaubt, für $0 < \xi_i < 1$ innerhalb von B zu liegen und für $\xi_i > 1$ sogar fehlklassifiziert zu werden. Um keine triviale Lösung zu erhalten, für die alle Schlupfvariablen groß sind, wird ein Parameter $C > 0$ eingeführt der $\xi_i > 0$ bestraft. Die Konstante C fungiert dabei als Tuning-Parameter. Da für den nicht trennbaren Fall keine optimal trennende Hyperebene existiert, sind Fehlklassifikationen, i. e. $y_i(x_i^T\beta + \beta_0) < 0$ mit $\xi_i > 1$ nicht zu vermeiden. Daher ist ein erster Ansatz ihre Anzahl durch ein C zu beschränken, indem das Optimierungsproblem um die Nebenbedingung $\sum_{i=1}^N \xi_i \leq C$ erweitert wird.
Eine weitere Möglichkeit ist es, den Term $C\sum_{i=1}^N \xi_i$ mit in das zu minimierende Funktional aufzunehmen. Für eine relative Abweichung von x_i von R um ξ_i nimmt man dafür eine Erhöhung des Zielfunktionals um $C\xi_i$ in Kauf. Je größer diese Rate C ist, desto stärker ist die Pönalisierung. Da für eine abgeschlossene, konvexe Menge $\mathcal{C} \subset \mathbb{R}^p$ gilt $\arg\min_{\beta\in\mathcal{C}} \frac{1}{2}||\beta||^2 = \arg\min_{\beta\in\mathcal{C}} ||\beta||$, lässt sich das Zielfunktional in (4.15) entsprechend umformulieren, ohne dabei die Lösung zu ändern. Das Optimierungsproblem zur Errechnung des sogenannten *C-Support-Vector-Klassifikators* (C-SVC) ergibt sich unter der Verwendung von $R = 1/||\beta||$ zu

$$\begin{aligned} &\min_{\xi,\beta_0,\beta} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^N \xi_i \\ &\text{unter den Nebenbedingungen } y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ &\text{und } \xi_i \geq 0 \text{ für alle } i = 1, \ldots, N. \end{aligned} \tag{4.16}$$

Eine Einführung in die Thematik wird z. B. von Hastie, Tibshirani und Friedman [7, S. 417-419], sowie von Schölkopf und Smola [15, S. 204-210] vorgenommen.

4.3.2 Berechnung des C-SVC

Da das zu minimierende Zielfunktional in (4.16) quadratisch ist und die Nebenbedingungen linear sind, lässt sich das Problem durch konvexe Optimierungstheorie lösen (s. Hastie, Tibshirani und Friedman [7, S. 420-421], sowie Schölkopf und Smola [15, S. 165-179]). Die zugehörige Lagrangefunktion des primären Problems

$$L_P(\xi, \beta_0, \beta, \alpha, \mu) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i(y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)) - \sum_{i=1}^N \mu_i\xi_i$$

ist bezüglich ξ, β_0 und β zu minimieren. Die Vektoren $\alpha = (\alpha_1, \ldots, \alpha_N)$ und $\mu = (\mu_1, \ldots, \mu_N)$ sind dabei die nichtnegativen Lagrange-Multiplikatoren. Durch Nullsetzen der entsprechenden partiellen Ableitung erhält man folgende Gleichungen zur Ermitt-

lung eines Sattelpunktes in (ξ, β_0, β)

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{4.17}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i, \tag{4.18}$$

$$\alpha_i = C - \mu_i, \tag{4.19}$$

sowie die Karush-Kuhn-Tucker-Bedingungen zur Ermittlung eines Sattelpunktes in (α, ξ)

$$\alpha_i(y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)) = 0 \tag{4.20}$$

$$\mu_i \xi_i = 0 \tag{4.21}$$

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) = 0 \tag{4.22}$$

$$\text{und } \alpha_i,\ \mu_i,\ \xi_i \geq 0 \ \text{ für alle } i = 1, \ldots, N. \tag{4.23}$$

Setzt man (4.17)-(4.19) in $L_P(\xi, \beta_0, \beta, \alpha, \mu)$ ein, so erhält man das duale Problem

$$\begin{aligned} &\text{Maximiere } L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_j^T x_i \\ &\text{unter den Nebenbedingungen } \sum_{i=1}^{N} \alpha_i y_i = 0 \\ &\text{und } 0 \leq \alpha_i \leq C \text{ für alle } i = 1, \ldots, N. \end{aligned} \tag{4.24}$$

Die Lösung $(\hat{\xi}, \hat{\beta}_0, \hat{\beta})$ ist durch die Gleichungen (4.17)-(4.23) eindeutig charakterisiert. Eine Lösung für β ist nach Gleichung (4.17) eine Linearkombination der Vektoren x_i und ergibt sich aus der Lösung $\hat{\alpha}$ des dualen Problems (4.24)

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i. \tag{4.25}$$

Bei der Errechnung von $\hat{\beta}$ haben nur die Beobachtungen i einen Einfluss, für die $\hat{\alpha}_i > 0$ gilt. Die entsprechenden Vektoren „stützen“ dabei die Hyperebene und bestimmen ihre Gestalt. Daher werden sie Stützvektoren (*support vectors*) genannt. Eine Lösung für μ erhält man aus (4.19)

$$\hat{\mu}_i = C - \hat{\alpha}_i,\ i = 1, \ldots, N.$$

Daher folgt für $0 \leq \hat{\alpha}_i < C$ mit (4.21), dass $\hat{\xi}_i = 0$ gilt und $\hat{\beta}_0$ kann in diesen Fällen durch (4.22) berechnet werden. Die übrigen Lösungen von ξ_i ergeben sich dann ebenfalls aus (4.22). Die durch $(\hat{\beta}_0, \hat{\beta})$ bestimmte Hyperebene führt zu dem C-Support-Vector-Klassifikator

$$\hat{G}(x) = \text{sign}(x^T\hat{\beta} + \hat{\beta}_0). \tag{4.26}$$

4.3.3 Nichtlineare Support-Vector-Klassifikatoren

Alle bisher genannten Klassifikationsmethoden, außer der quadratischen Diskriminanzanalyse, gehen von einer linearen Entscheidungsgrenze zwischen zwei Gruppen aus. Diese Annahme ist jedoch nicht für alle Anwendungen adäquat. Die Verallgemeinerung der SVM-Methode basiert darauf, die Eingabewerte aus dem Raum $\mathcal{X}$ durch eine Transformation

$$\Phi : \mathcal{X} \to \mathcal{H}$$

in einen höherdimensionalen Raum $\mathcal{H}$ zu überführen (s. Hastie, Tibshirani und Friedman [7, S. 423-426], sowie Schölkopf und Smola [15, S. 25-60,S. 200-204]). Dabei seien hier zur Vereinfachung $\mathcal{X}$ und $\mathcal{H}$ endlich-dimensionale, reelle Vektorräume ausgestattet mit dem Standardskalarprodukt $\langle \cdot, \cdot \rangle$. Man erhofft sich dadurch, den Trainingsdatensatz $(\Phi(x_i), y_i),\ i = 1, \ldots, N$ in $\mathcal{H}$ besser linear trennen zu können. Die rücktransformierte Entscheidungsgrenze

$$\Phi^{-1}(H) = \{x \in \mathcal{X} : \hat{\beta}^T \Phi(x) + \hat{\beta}_0 = 0\}$$

ist im Allgemeinen nichtlinear. Transformiert man zum Beispiel die Daten durch

$$\Phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3,\ (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2),$$
$$\Phi^{-1}(H) = \{x \in \mathbb{R}^2 : \hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_2^2 + \hat{\beta}_3 \sqrt{2} x_1 x_2 + \hat{\beta}_0 = 0\},$$

so wird die rücktransformierte Hyperebene durch eine quadratische Form beschrieben. Die Komponentenfunktionen von Φ heißen in diesem Zusammenhang auch Basisfunktionen, da die Entscheidungsgrenze mit Hilfe einer Linearkombination von ihnen beschrieben wird. Die *Kernelfunktion* bezüglich Φ ist definiert durch

$$\mathcal{K}(x, x') = \langle \Phi(x), \Phi(x') \rangle\ ,\ x, x' \in \mathcal{X}. \tag{4.27}$$

Eine Kernelfunktion ist symmetrisch und positiv (semi-)definit. Um das entsprechende duale Problem (4.24) zu lösen, ist nur die Kenntnis über die zugrundeliegende Kernelfunktion notwendig, da die Lagrange-duale Funktion gegeben ist durch

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathcal{K}(x_i, x_j).$$

Die Lösung $\hat{\beta}$ ist hier analog zu (4.25) eine Linearkombination der transformierten Vektoren $\Phi(x_i)$.

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \Phi(x_i)$$

Für den Fall, dass $0 \leq \hat{\alpha}_i < C$ gilt, errechnet sich $\hat{\beta}_0$ gemäß

$$\hat{\beta}_0 = -\langle \hat{\beta}, x_i \rangle = -\sum_{j=1}^{N} \hat{\alpha}_j y_j \mathcal{K}(x_j, x_i).$$

Um die Hyperebene in dem höher dimensionalen Raum durch die Gleichung (4.28) zu bestimmen, ist also nur die Festlegung einer Kernelfunktion und die Lösung des dualen Problems notwendig.

$$H = \left\{ x \in H : \langle \hat{\beta}, \Phi(x) \rangle + \hat{\beta}_0 = \sum_{j=1}^{N} \hat{\alpha}_j y_j \mathcal{K}(x_j, x) + \hat{\beta}_0 = 0 \right\} \tag{4.28}$$

Die Berechnungen in Kapitel 4.3.2 zur Bestimmung einer Hyperebene sind dabei für die Kernelfunktion $\mathcal{K}(x, x') = \langle x, x' \rangle$ formuliert. Tauscht man die Kernelfunktion in einem Algorithmus durch eine andere aus, so spricht man von dem *Kernel-Trick*. Legt man nun zuerst eine symmetrische, positiv semidefinite Kernelfunktion fest, so ist dabei die Existenz einer Transformation Φ mit (4.27) theoretisch durch den Satz von Mercer [15, S. 37-38] garantiert. Im Zusammenhang mit der Support-Vector-Machine können zum Beispiel folgende Kernelfunktionen gewählt werden:

$$\begin{aligned} &\text{Polynomialer Kernel: } \mathcal{K}(x, x') = \langle x, x' \rangle^d, \ d \in \mathbb{N}, \\ &\text{Radiale-Basisfunktionen-Kernel: } \mathcal{K}(x, x') = \exp(-\langle x, x' \rangle / \sigma^2), \ \sigma^2 > 0, \\ &\text{Neuronales Netz: } \mathcal{K}(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2), \ \kappa_1 > 0, \ \kappa_2 \in \mathbb{R}. \end{aligned} \tag{4.29}$$

4.4 Klassifikations- und Regressionsbäume (CART)

Die Methode der Klassifikations- und Regressionsbäume (CART: *classification and regression trees*) zur Klassifizierung von Daten wurde von Breiman [4] entwickelt. Er verdeutlichte dabei, dass Bäume auch außerhalb der Modellierung von Datenstrukturen in der Informatik für statistische Anwendungen relevant sind. Im Folgenden wird zunächst der Begriff des Baums als Objekt der Graphentheorie nach Diestel [5, S. 1-15] eingeführt.

4.4.1 Bäume als Klassifikatoren

Ein ungerichteter Graph G ist ein Tupel $G = (V, E)$ disjunkter Mengen mit $E \subset V \times V$, dabei heißen die Elemente von V Knoten, die Elemente von E Kanten. Im Folgenden sei V nichtleer und E sowie V endlich. Für eine Kante $\{x, y\} \in E$ schreibt man kürzer xy oder auch yx. Sei $e \in E$, so beschreibt man durch $G - e = (V, E \setminus \{e\})$ den Graphen, der durch Löschung der Kante e entsteht und analog $G + e = (V, E \cup \{e\})$ für $e \in V \times V$. Der Graph G' heißt Teilgraph von G, wenn $V' \subset V$ und $E' \subset E$ gilt. Ein Weg von x_0 nach x_m ist ein Graph $P = (V, E)$ mit $V = \{x_0, x_1, \ldots, x_m\}$ und $E = \{x_0x_1, \ldots, x_{m-1}x_m\}$. Man schreibt kurz: $P = x_0x_1 \ldots x_m$. Mit Hilfe der Menge aller Kanten, die den Knoten v beinhalten $E(v) = \{e \in E : v \in e\}$, lässt sich der Grad einer Ecke durch $d(v) = |E(v)|$ definieren. Wenn für alle $x, y \in E$ ein Weg von x nach y existiert, so heißt ein Graph zusammenhängend. Definiert man nun die Äquivalenzrelation $x \sim y$, falls ein Weg von x nach y existiert, so heißen die Äquivalenzklassen Zusammenhangskomponenten. Ein Kreis ist ein Weg mit $P = x_0x_1 \ldots x_{m-1}x_0$ für $m \geq 3$.
Ein *Wald* ist ein Graph, der keinen Kreis als Teilgraphen enthält. Ein zusammenhängender Wald ist ein *Baum*. Somit ist ein Wald ein Graph, dessen einzelne Zusammenhangs-

komponenten Bäume sind. Bäume können nach Diestel [5, S. 13] alternativ definiert werden:

Satz 4.2. Die folgenden Aussagen sind äquivalent für einen Graphen T:

1. T ist ein Baum;
2. zwischen je zwei Ecken enthält T genau einen Weg;
3. T ist minimal zusammenhängend, d.h. T ist zusammenhängend aber für jede Kante e von T ist $T - e$ nicht zusammenhängend;
4. T ist maximal kreislos, d.h. T enthält keinen Kreis aber für je zwei Ecken x, y mit $xy \notin E$ enthält $T + xy$ einen Kreis.

Aus diesem Satz folgt insbesondere, dass ein zusammenhängender Graph mit m Ecken genau dann ein Baum ist, wenn er $m - 1$ Kanten hat. Die Knoten t eines Baumes mit $d(t) = 1$ heißen *Blätter*, alle anderen *innere Knoten*. Mit $B(T)$ sei die Menge aller Blätter eines Baums T bezeichnet.
Wählt man ein $t_1 \in T$, so lässt sich eine Halbordnung definieren durch $x \leq y$ falls ein Weg P existiert mit $P = t_1 \dots x \dots y$. Der Knoten t_1 heißt in diesem Zusammenhang dann *Wurzel* und T *Wurzelbaum*. Diese Halbordnung ermöglicht eine Modellierung hierarchischer Datenstrukturen. Gilt für $x \leq y$ zusätzlich $xy \in E$, so heißt x ein *Vaterknoten* von y und y ein *Kindknoten* von x. Ein *Binärbaum* ist ein Wurzelbaum, dessen innere Knoten je ein oder zwei Kindknoten besitzen.

In der Situation von Kapitel 2 ist ein Klassifikationsbaum frei nach Breiman [4, S. 20-23] wie folgt definiert.

Definition 4.3. Ein Graph heißt *Klassifikationsbaum* $T = (V, E)$ mit der Zuordnungsfunktion κ, wenn $T = (V, E)$ ein Binärbaum mit $V = \{t_1, \dots, t_M\}$ und der Wurzel t_1 ist, für den jeder innere Knoten genau zwei Kindknoten besitzt und wenn die Funktion $\kappa : V \to \mathcal{P}(\mathcal{X}),\ t \longmapsto \kappa(t) = \mathcal{X}_t$ folgende Eigenschaften erfüllt:

1. $\kappa(t_1) = \mathcal{X}$;
2. Für jeden inneren Knoten t und seine zwei Kindknoten l und r gilt
$$\mathcal{X}_t = \mathcal{X}_l \dot{\cup} \mathcal{X}_r;$$
3. Triviale Partitionen sind ausgeschlossen: $\mathcal{X}_t \neq \emptyset\ \forall t \in T$.

Jeder innerer Knoten t steht so für eine Teilung der Menge $\mathcal{X}_t$ in zwei disjunkte Teilmengen $\mathcal{X}_l$ und $\mathcal{X}_r$. Aufgrund dieser sukzessiven Teilung des Eingaberaums gilt

$$\mathcal{X} = \dot{\bigcup}_{t \in B(T)} \mathcal{X}_t.$$

Liegt ein Trainingsdatensatz vor, so beschreibt $N_t = |\{x_i : x_i \in \mathcal{X}_t\}|$ die Anzahl aller Beobachtungen, die in $\mathcal{X}_t$ liegen. Der relative Anteil der zur Gruppe k gehörenden Beobachtungen in $\mathcal{X}_t$ ist dann gegeben durch

$$\hat{p}_{kt} = \frac{1}{N_t} \sum_{i:x_i \in \mathcal{X}_t} I(g_i = k).$$

Wird ein neuer Eingabewert beobachtet, so entscheidet sich der Klassifikationsbaum mittels

$$\hat{G}(x) = \arg\max_{k \in \mathcal{G}} \sum_{t \in B(T)} \hat{p}_{kt} I(x \in \mathcal{X}_t) = \sum_{t \in B(T)} k(t) I(x \in \mathcal{X}_t) \tag{4.30}$$

für die Gruppe $k\,(t) = \arg\max_{k \in \mathcal{G}} \hat{p}_{kt}$, die am häufigsten vertreten ist in der Teilmenge des Eingaberaums $\mathcal{X}_t$ in der x liegt. Die A-posteriori-Wahrscheinlichkeiten werden hier durch $\hat{p}_k(x) = \sum_{t \in B(T)} \hat{p}_{kt} I(x \in \mathcal{X}_t)$ geschätzt.

4.4.2 Konstruktionsalgorithmus

Es stellt sich die Frage, nach welchen Kriterien die Partitionen gewählt werden, sollen um einen Klassifikationsbaum zu konstruieren. Für ein Blatt t lässt sich die mit ihm assoziierte Menge $\mathcal{X}_t \subset \mathbb{R}^p$ teilen in

$$\mathcal{X}_l\,(j, \beta) = \left\{ x = (x_1, \ldots, x_p)^T \in \mathcal{X}_t : x_j \leq \beta \right\} \text{ und } \mathcal{X}_r\,(j, \beta) = \{x \in \mathcal{X}_t : x_j > \beta\}\,.$$

Eine *Teilung* (*split*) ist also durch die Festlegung, in welcher Dimension j und zu welchem Wert β die jeweilige Menge $\mathcal{X}_t$ entzweit wird, bestimmt. Die Menge aller nichttrivialen Teilungen eines Blatts t sei gegeben durch

$$\mathcal{S}_t = \{(j, \beta) : j \in \{1, \ldots, p\},\ \beta \in \mathbb{R},\ \mathcal{X}_l(j, \beta) \neq \emptyset \neq \mathcal{X}_r(j, \beta)\}.$$

Verlustfunktionen wie in Kapitel 3.1 ermöglichen es, Fehlklassifikationen zu quantifizieren um so verschiedene Entscheidungsregeln miteinander vergleichen zu können. Analog führt Breiman den Begriff des *Unreinheitsmaßes* (*impurity measure*) $q : V \to \mathbb{R}_{\geq 0}$ ein, um Knoten bezüglich ihrer Klassifikationsregel $k(t)$ zu vergleichen. In der Literatur werden zur Wahl von q verschiedene Vorschläge gemacht.

$$\text{Anteil an Fehlklassifikationen: } q(t) = \frac{1}{N_t} \sum_{i:x_i \in \mathcal{X}_t} I(g_i \neq k(t)) = 1 - \hat{p}_{k(t)t}$$

$$\text{Gini-Index: } q(t) = \sum_{k \neq k'} \hat{p}_{kt} \hat{p}_{k't} = \sum_{k=1}^{K} \hat{p}_{kt}\,(1 - \hat{p}_{kt})$$

$$\text{Kreuzentropie: } q(t) = -\sum_{k=1}^{K} \hat{p}_{kt} \log \hat{p}_{kt}$$

Der Gini-Index lässt sich dabei folgendermaßen interpretieren. Teilt man an einem Knoten eine Beobachtung randomisiert mit der Wahrscheinlichkeit $\hat{p}_{kt}$ in die Gruppe k ein,

$$\mathbb{P}(\hat{G}(x) = k) = \hat{p}_{kt}, \text{ für } x \in \mathcal{X}_t,$$

so ist der erwartete Trainingsfehler gerade der Gini-Index

$$\begin{aligned} \mathbb{E}\frac{1}{N_t}\sum_{i:x_i\in\mathcal{X}_t} I(g_i \neq \hat{G}(x)) &= \frac{1}{N_t}\sum_{i:x_i\in\mathcal{X}_t} \mathbb{P}(g_i \neq \hat{G}(x)) \\ &= \sum_{k=1}^{K}\frac{1}{N_t}\sum_{i:x_i\in\mathcal{X}_t} \mathbb{P}(\hat{G}(x) \neq k) I(g_i = k) \\ &= \sum_{k=1}^{K}\hat{p}_{kt}(1-\hat{p}_{kt}). \end{aligned} \tag{4.31}$$

Teilt man nun das Blatt t, bzw. die Partition $\mathcal{X}_t$, so gibt die *Güte einer Teilung* (*goodness of split*) $s \in \mathcal{S}_t$ die Differenz an, um die sich die Schätzung verbessert:

$$\triangle q(s,t) = q(t) - \frac{N_{l(s)}}{N_t}q(l(s)) - \frac{N_{r(s)}}{N_t}q(r(s)).$$

Für jedes Blatt kann so eine optimale Teilung gefunden werden.

$$\hat{s}_t = \arg\max_{s\in\mathcal{S}_t} \triangle q(s,t) \tag{4.32}$$

Die zu maximierende Gütefunktion verändert sich dabei nur an endlich vielen Stellen. In einem zweiten Schritt wird sich für die Teilung des Blattes entschieden, die die Schätzung bezüglich des gewählten Unreinheitsmaßes am stärksten verbessert.

$$\hat{t} = \arg\max_{t\in B(T)} \triangle q(\hat{s}_t, t) \tag{4.33}$$

Der Baum wird dann um die entsprechenden Knoten und Kanten erweitert.

$$\tilde{T} = (V \cup \{l(\hat{s}_{\hat{t}}), r(\hat{s}_{\hat{t}})\}, E \cup \{tl(\hat{s}_{\hat{t}}), tr(\hat{s}_{\hat{t}})\}) \tag{4.34}$$

Die so entstehenden Bäume werden so lange nach (4.32)-(4.34) erweitert, bis ein Stopp-Kriterium den Konstruktionsalgorithmus abbricht:

1. Teile die Blätter solange die Anzahl der Beobachtungen in jedem Blatt größer als $C \in \mathbb{N}$ ist;

2. Spalte den Knoten t nur, wenn die Güte seiner optimalen Teilung größer als eine Rate $C > 0$ ist.

Zusammengefasst wird ein Klassifikationsbaum folgendermaßen konstruiert:

Algorithmus 4.4.
Input: Trainingsdatensatz $(x_i, g_i),\ i = 1, \ldots, N$. Setze $T = (\{t_1\}, \emptyset)$ mit $\mathcal{X}_{t_1} = \mathcal{X}$.
Wiederhole, solange das Stopp-Kriterium nicht erfüllt ist, folgende Schritte:

1. Finde für jedes Blatt $t \in T$ die optimale Teilung nach (4.32).
2. Finde die optimale Teilung nach (4.33).
3. Bestimme $\tilde{T}$ nach (4.34).
4. Setze $T \leftarrow \tilde{T}$.

Output: T.

Ausgehend von dem Unreinheitsmaß eines Knoten lässt sich das *Unreinheitmaß eines Baums* (*tree impurity*) definieren als

$$Q(T) = \sum_{t \in B(T)} \frac{N_t}{N} q(t). \tag{4.35}$$

Wählt man dabei das Unreinheitsmaß eines Knotens als den Anteil an Fehlklassifikationen, so ist das Unreinheitsmaß des Baumes gerade der Trainingsfehler

$$\begin{aligned} \mathrm{err}(\hat{G}) &= \frac{1}{N} \sum_{i=1}^{N} I(g_i \neq \hat{G}(x_i)) = \frac{1}{N} \sum_{t \in B(T)} \sum_{i: x_i \in \mathcal{X}_t} I(g_i \neq k(t)) \\ &= \frac{1}{N} \sum_{t \in B(T)} (N_t - \sum_{i: x_i \in \mathcal{X}_t} I(g_i = k(t))) = \frac{1}{N} \sum_{t \in B(T)} N_t - N_t \hat{p}_{k(t)t} \\ &= \sum_{t \in B(T)} \frac{N_t}{N} q(t) = Q(T). \end{aligned} \tag{4.36}$$

Sei $\tilde{T}$ wie in (4.34), so gilt

$$\begin{aligned} \triangle Q(s, T) = Q(T) - Q(\tilde{T}) &= \sum_{t' \in B(T)} \frac{N_{t'}}{N} q(t') - \sum_{t' \in B(\tilde{T})} \frac{N_{t'}}{N} q(t') \\ &= \frac{N_t}{N} q(t) - \frac{N_l}{N} q(l) - \frac{N_r}{N} q(r) = \frac{N_t}{N} (q(t) - \frac{N_l}{N_t} q(l) - \frac{N_r}{N_t} q(r)) \\ &= \frac{N_t}{N} \triangle q(s, t). \end{aligned} \tag{4.37}$$

Dieses Verfahren führt jedoch zu einer hohen Anzahl an Knoten des resultierenden Entscheidungsbaumes T und damit zu einer Überanpassung an den Trainingsdatensatz (s. Breiman [4], sowie Hastie, Tibshirani und Friedman [7, S. 307-310]).

4.4.3 Tree-Pruning

In einem zweiten Schritt, dem *tree pruning* (Baumschnitt), wird der Klassifikationsbaum wieder verkleinert, um die Komplexität des Modells adäquat zu wählen. Für jedes $\alpha \geq 0$ ist ein Teilbaum $T_\alpha \subset T$ zu finden mit

$$T_\alpha = \arg\min_{\tilde{T} \subset T} \sum_{t \in B(\tilde{T})} \frac{N_t}{N} q(t) + \alpha |B(\tilde{T})|. \tag{4.38}$$

Da nur endlich viele Teilbäume existieren, stehen auch nur endlich viele Werte für α zur Auswahl. Der Tuning-Parameter α wird mit Hilfe einer n-fachen Kreuzvalidierung geschätzt. Dabei wird der Datensatz in n gleichgroße Datenmengen geteilt. Für jedes $j = 1, \ldots, n$ wird durch Auslassung der j-ten Datenteilmenge ein Klassifikationsbaum T^j konstruiert und durch Pruning das α_j mit dem Teilbaum $T^j_{\alpha_j}$ bestimmt, das (4.38) minimiert. Letztendlich wird mit dem Schätzer $\hat{\alpha} = \frac{1}{n} \sum_{j=1}^{n} \alpha_j$ der Klassifikationsbaum $T_{\hat{\alpha}}$ als Ergebnis gewählt und damit eine Klassifikationsregel gemäß (4.30). Dieses Verfahren wird von James et al. [9, S. 306-312] und Hastie, Tibshirani und Friedman [7, S. 307-309] erläutert.

5 Ensemble-Methoden

5.1 Bootstrapping und Bagging

Das *Bootstrapping*, die Generierung von Datensätzen auf Grundlage der beobachteten Stichproben, ermöglicht eine Schätzung des (erwarteten) Testfehlers der Klassifikationsmethoden. Dabei werden n zufällige Datensätze (x_i, g_i) aus dem beobachteten Trainingsdatensatz Z ausgewählt. Da diese Stichprobenziehung mit Zurücklegen erfolgt, kann ein Bootstrap-Datensatz Z^{*b} dieselben Daten mehrfach enthalten. Wiederholt man diese Ziehung B mal und passt für jedes Z^{*b} ein Modell mit Klassifikationsregel $\hat{G}^{*b}$ an, so wäre es ein erster Ansatz, den erwarteten Testfehler durch

$$\widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(g_i, \hat{G}^{*b}(x_i))$$

zu schätzen. Da die Bootstrap-Datensätze aus den beobachteten Daten des Trainingsdatensatzes bestehen, aus dem wiederum die Stichproben zur Berechnung stammen, wird der erwartete Testfehler durch eine Überanpassung von $\widehat{\mathrm{Err}}_{\mathrm{boot}}$ systematisch unterschätzt. Motiviert durch das Kreuzvalidierungsverfahren kann die Schätzung verbessert werden durch

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(g_i, \hat{G}^{*b}(x_i),$$

wobei C^{-i} die Menge an Indizes der Bootstrap-Datensätze ist, die die i-te Beobachtung nicht enthalten. Dabei sei $|C^{-i}| \neq 0$ für alle i vorausgesetzt. Das durch das Kreuzvalidierungsverfahren entstehende Bias lässt sich durch eine Modifizierung durch den *.632-Schätzer*

$$\widehat{\mathrm{Err}}^{(.632)} = 0.368 \cdot \mathrm{err} + 0.632 \cdot \widehat{\mathrm{Err}}^{(1)}$$

korrigieren. Eine Ausführliche Herleitung des .632-Schätzers wird von Hastie, Tibshirani und Friedman [7, S. 250-251] gegeben.
Die Idee des *Baggings*, kurz für *bootstrap aggregation*, ist es die Varianz durch Kombination mehrerer Klassifikatoren zu reduzieren. Für kategoriale Daten ist das mit dem Modus, dem am häufigsten auftretenden Wert gemäß

$$\hat{G}_{\mathrm{bag}}(x) = \mathrm{modus}(\hat{G}^{*1}(x), \ldots, \hat{G}^{*B}(x)) \tag{5.1}$$

zu erreichen. Die Klassifikationsregel ordnet die Beobachtung in die Gruppe ein, in der es von der Mehrheit aller anderen Regeln eingeordnet wird (s. Hastie, Tibshirani und Friedman [7, S. 282-283], sowie James et al. [9, S. 187-190]).

5.2 Random Forests

Breimans Idee [3] der zufälligen Wälder (*Random Forests*) basiert auf der Konstruktion von Klassifikationsbäumen nach zufälligen Gesichtspunkten. Dabei kommt für die Teilung jedes Knoten nur eine beliebig gewählte Teilmenge der Eingabewerte in Frage, während keine Reduzierung der Bäume durch Pruning stattfindet.

Algorithmus 5.1.

1. Für $b = 1, \ldots, B$:
   Wiederhole folgende Schritte zur rekursiven Konstruktion eines Baumes mit Klassifikationsregel $\hat{G}_b(x)$ bis er eine Mindestanzahl an Blättern M_{min} besitzt. Starte dabei mit dem Baum der nur einen Knoten besitzt.
   a) Wähle zufällig m der verschiedenen p Eingabewerte aus.
   b) Überprüfe das i-te Blatt darauf, welche Halbierung gemäß (4.32) dieser m Variablen eine gegebene Verlustfunktion minimiert.
   c) Teile das Blatt, das dies am besten bezüglich einer minimalen Verlustfunktion tut.
2. Output:
$$\hat{G}_{rf}(x) = \text{modus}(\hat{G}_1(x), \ldots, \hat{G}_B(x)).$$

Der Algorithmus wird in dieser Form von Hastie, Tibshirani und Friedman [7, S. 587-588] wiedergegeben.

5.3 Boosting

Ein wichtiges Merkmal von Klassifikationsbäumen ist, dass sich die Gewichtung und damit die Relevanz einer Beobachtung verändern lässt. Durch die Beobachtungsgewichte w_i mit $w_i \geq 0$ und $\sum_{i=1}^{N} w_i = 1$ lassen sich bestimmte Beobachtungen priorisieren.

$$N_t = N \cdot \sum_{i=1}^{N} w_i I(x_i \in \mathcal{X}_t)$$

Mit Hilfe von Boosting-Algorithmen ist es möglich, den Trainingsfehler einer Methode ähnlich wie beim Bagging durch ein Ensemble an Klassifikationsregeln zu reduzieren. Dabei zählt jedoch nicht jede Stimme der zugrundeliegenden Klassifikatoren gleich, sondern wird entsprechend seiner Vorhersageleistung gewichtet.
Der in diesem Kapitel vorgestellte Boosting-Algorithmus SAMME von Zhu et al. [18] verwendet zur Minimierung des Klassifikationsfehlers eine multinomiale Verallgemeinerung der exponentiellen Verlustfunktion

$$L(y, f) = \exp\left(-\frac{1}{K}\left(\sum_{k=1}^{K} y_k f_k\right)\right) = \exp\left(-\frac{1}{K} y^T f\right)$$

mit $y = (y_1, \ldots, y_K)$, $f = (f_1, \ldots, f_K) \in \mathbb{R}^K$. Der Ausgabewert wurde dabei codiert gemäß

$$y_k = \begin{cases} 1, & \text{falls } g_i = k, \\ -\frac{1}{K-1}, & \text{falls } g_i \neq k. \end{cases}$$

Für f soll zusätzlich gelten

$$\sum_{k=1}^{K} f_k = 0 \text{ und } f_k = \begin{cases} 1, & \text{für genau ein } k, \\ -\frac{1}{K-1}, & \text{sonst.} \end{cases} \tag{5.2}$$

Für jede Funktion $f(x)$, für die (5.2) für alle $x \in \mathbb{R}^p$ gilt, existiert genau eine Klassifikationsregel mit $\hat{G}(x) = k$, falls $f_k(x) = 1$ und umgekehrt existiert für jede Klassifikationsregel eine Funktion mit

$$f_k(x) = \begin{cases} 1, & \text{falls } \hat{G}(x) = k, \\ -\frac{1}{K-1}, & \text{falls } \hat{G}(x) \neq k. \end{cases}$$

Gesucht ist die Funktion, deren erwarteter Verlust bezüglich der exponentiellen Verlustfunktion minimal ist.

$$f(x) = \arg\min_f \mathbb{E}_{Y|X=x} \exp\left(-\frac{1}{K}\left(\sum_{k=1}^{K} Y_k f_k(x)\right)\right) \text{ unter } \sum_{k=1}^{K} f_k(x) \tag{5.3}$$

Setzt man f als Linearkombination von Basisfunktionen g_m voraus

$$f(x) = \sum_{m=1}^{M} \alpha_m g_m(x), \ \alpha_m \in \mathbb{R},$$

so ist die revolutionär neue Idee des Boostings, eine Lösung von (5.3) stufenweise in M Schritten zu approximieren. Für jede Basisfunktion gelte dabei Eigenschaft (5.2).

Algorithmus 5.2 (Forward Stagewise Additive Modeling)**.**

1. Setze $f_0(x) = 0$.
2. Für $m = 1$ bis M:
   a) Berechne
   $$(\alpha_m, g_m) = \arg\min_{\alpha, g} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \alpha g(x_i)) \text{ unter (5.2).}$$
   b) Setze $f_m(x) = f_{m-1}(x) + \alpha_m g_m(x)$.
3. Output: $\hat{G}_{boost}(x) = \arg\min_{k \in \mathcal{G}} f_{M,k}(x)$.

Bei der Konstruktion des m-ten Klassifikators lernt dieser von seinen Vorgängern und wird nach seiner Vorhersageleistung α_m gewichtet. Verwendet man die exponentielle Verlustfunktion, so ergibt sich das Minimierungsproblem mit den Beobachtungsgewichten $w_i = \exp\left(-\frac{1}{K} y_i^T f_{m-1}(x_i)\right)$ zu

$$\begin{aligned} (\alpha_m, g_m) &= \arg\min_{\alpha, g} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \alpha g(x_i)) \\ &= \arg\min_{\alpha, g} \sum_{i=1}^{N} w_i \exp\left(-\frac{1}{K} \alpha y_i^T g(x_i)\right). \end{aligned} \tag{5.4}$$

Anhand der Lösung von (5.4), gegeben durch

$$\hat{G}_m(x) = \arg\min \sum_{i=1}^{N} w_i I(g_i \neq \hat{G}(x_i)), \quad \alpha_m = \frac{(K-1)^2}{K}\left(\log\frac{1-\mathrm{err}_m}{\mathrm{err}_m} + \log(K-1)\right),$$

mit dem gewichteten Trainingsfehler $\mathrm{err}_m = \sum_{i=1}^{N} w_i I(y_i \neq \hat{G}_m(x_i)) / \sum_{i=1}^{N} w_i$, zeigt Zhu et al. [18] das der Algorithmus 5.3 dieselbe Klassifikationsregel ergibt. So verdient er seinen Namen SAMME (*forward stagewise additive model using a multi-class exponential loss function*).

Algorithmus 5.3 (SAMME)**.**

1. Input: Trainingsdatensatz $(x_i, g_i),\ i = 1, \ldots, N$.
2. Setze die Beobachtungsgewichte $w_i = 1/N,\ i = 1, \ldots, N$.
3. Für $m = 1$ bis M:
   a) Trainiere einen CART-Klassifikator $\hat{G}_m(x)$ mit den Gewichten w_i.
   b) Berechne den gewichteten Trainingsfehler
   $$\mathrm{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq \hat{G}_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   c) Berechne
   $$\alpha_m = \log\frac{1-\mathrm{err}_m}{\mathrm{err}_m} + \log(K-1).$$
   d) Berechne für $i = 1, \ldots, N$:
   $$\tilde{w}_i = w_i \exp(\alpha_m I(c_i \neq \hat{G}_m(x_i))).$$
   e) Setze neue Beobachtungsgewichte für i=1,. . . ,N:
   $$w_i \leftarrow \frac{\tilde{w}_i}{\sum_{j=1}^{N} \tilde{w}_j}.$$
4. Output: $\hat{G}_{boost}(x) = \arg\max_k \sum_{m=1}^{M} \alpha_m I\left(\hat{G}_m(x) = k\right)$.

Sinnvoll ist die Konstruktion des SAMME-Klassifikators nur für positive Gewichte α_m. Setzt man $1 - \frac{1}{K} > \mathrm{err}_m$ voraus, so lässt sich zeigen, dass diese positiv sind.

6 Medizinische Grundlagen

6.1 Bronchoalveoläre Lavage (BAL) und seine Parameter

Ein wichtiges Hilfsmittel bei der Diagnose von Krankheiten ist die Bestimmung der Anzahl der Zellen der Immunabwehr. Eine erhöhte Anzahl deutet auf ein sich wehrendes Immunsystem und damit auf das Vorhandensein von Krankheitserregern hin. So kann zum Beispiel die *bronchoalveoläre Lavage* (BAL), eine bestimmte Lungenspülung und eine anschließende Auswertung der Zellenanzahl, die Diagnose von Lungenkrankheiten erleichtern. Sie heißt bronchoalveolär, weil sie aus den Luftröhrenästen (Bronchien) und Lungenbläschen (Alveolen) gewonnen wird (s. Kroegel und Bonella [11, S. 144]).
Für ein besseres Verständnis immunbiologischer Grundbegriffe werden diese zunächst näher erläutert (s. Murphy, Travers und Walport [13, S. 4 und S. 7-14]). Die Zellen des Immunsystems entwickeln sich allesamt aus Vorläuferzellen aus dem Rückenmark, den hämatopoetischen Stammzellen. Die dabei entstehenden Zellen können in zwei Hauptgruppen eingeteilt werden, den myeloiden und den lymphatischen Zellen. Die entwickelten Zellen des Immunsystems wandern in peripheres Gewebe, kommen so auch in die Lunge, zirkulieren durch die Blutbahnen oder durch ein darauf spezialisiertes Gefäßsystem, dem lymphatischen System.
Aus den gemeinsamen myeloiden Vorläuferzellen gehen die *Makrophagen* (Fresszellen), *Granulozyten*, *Mastozyten* (Mastzellen) und roten Blutkörperchen hervor. Makrophagen tragen entscheidend zur Entstehung von Entzündungen bei, die für eine erfolgreiche Immunreaktion unerlässlich sind und können eindringende Mikroorganismen in sich aufnehmen und töten. Granulozyten lassen sich wiederum in *neutrophile*, *eosinophile* und *basophile* Granulozyten unterscheiden. Antwortet das Immunsystem auf eine Infektion, werden sie vermehrt produziert und wandern zur infizierten Stelle. Neutrophile nehmen verschiedene Mikroorganismen auf und zerstören diese, unter anderem durch in ihnen gespeicherte abbauende Enzyme. Die genaue Bedeutung der Eosinophile und Basophile für das Immunsystem ist nicht genau bekannt, doch weiß man, dass sie an allergischen Entzündungsreaktionen beteiligt sind. Mastzellen sind ebenfalls an allergischen Reaktionen beteiligt und unterstützen das Auslösen von Entzündungen.
Die bisher erwähnten Zellen gehören zum sogenannten angeborenen Immunsystem. Sie können eine Vielzahl unterschiedlicher Krankheitserreger bekämpfen, jedoch keine dauerhafte Immunität gegen bestimmte Erreger erreichen. Eine wichtige Gruppe von *Lymphozyten* (weiße Blutkörperchen), die aus den lymphatischen Vorläuferzellen hervorgehen, sind die antigenspezifischen Lymphozyten. Indem ihre Rezeptoren durch Antigene aktiviert werden, zum Beispiel durch eine Impfung, ist eine Immunität gegenüber bestimmten Krankheiten möglich. Sie gehören daher zum adaptiven Immunsystem.
Bei der BAL wird eine aus mindestens 100 ml bestehende, 0.9-%ige NaCl-Lösung in verschiedene Segmente der Lunge gespült, um bronchoalveoläre Zellen für eine Differential- und Immunzytologie zu gewinnen. Sie dient zur Identifizierung interstitieller Lungenkrankheiten, siehe Kapitel 6.3 und der Sarkoidose. Bei der Differentialzytologie wird die Verteilung der Zellen des Immunssystems betrachtet. Ist sie unauffällig, liegt also ein Normalbefund nach Tabelle 4 vor, ist eine interstitielle Erkrankung unwahrscheinlich.

Eine genauere Betrachtung der verschiedenen Lymphozyten-Typen in Form einer Immunzytologie kann weitere Indizien für eine Diagnose ergeben. Die für die Untersuchung zurückgewonnene Lavageflüssigkeit (BAL-Recovery) sollte mehr als 30% der ursprünglich instillierten 100 ml betragen und die BAL sollte unverfälscht von Blut sein, sonst ist das Ergebnis nicht aussagekräftig genug. Wird zu wenig Lavageflüssigkeit zurückgewonnen, dominieren die Zellen der Bronchien, vorallem Neutrophile und Makrophagen, die BAL, da die Lavage nicht in ausreichendem Maß zu den Alveolen durchdringt. Die Anzahl der Zellen in der BAL ist dabei in 10^6 Zellen pro Milliliter angegeben (s. Kroegel und Bonella [11, S. 144-146]).

Parameter	Sollwert	Pathologisch	Indiz für
Zellenanzahl	$3 \cdot 10^6$ bis $10 \cdot 10^6$/ml	$< 3 \cdot 10^6$/ml $> 10 \cdot 10^6$/ml	BAL-Recovery < 30 ml EAA, Pneumonie
Lymphozyten	$\leq 18\%$	Lymphozytose ($> 40\%$)	EAA, aktive Sarkoidose
Neutrophile	$\leq 3\%$	Neutrophilie ($> 40\%$)	bakterielle Pneumonien
Eosinophile	$\leq 1\%$	Eosinophilie ($> 20\%$)	eosinophilenassoziierte Erkrankung

Tabelle 4: Differentialzytologie [11, Tab. 2.119 u. Tab. 2.120, S. 146-147]

6.2 Spirometrische Parameter und Packungsjahr

6.2.1 Inspiratorische Vitalkapazität (IVC)

Ein wichtiger Bestandteil einer pneumologischen Untersuchung ist die Messung der Lungenfunktionswerte mit Hilfe von Spirometern. Die Vitalkapazität (VC) ist der elementarste Begriff der Spirometrie und ist definiert als die Differenz des Lungenvolumens zwischen der maximalen Ein- und Ausatmung. Zur Ermittlung der *inspiratorischen Vitalkapazität* (IVC) wird zuerst maximal aus- und dann eingeatmet, für die expiratorische VC zuerst ein- und dann langsam ausgeatmet. Bei korrekter Messung sind die beiden Größen gleich groß. Wird die Vitalkapazität berechnet, indem die Aus- bzw. Einatmung so schnell wie möglich erfolgt, spricht man auch von der forcierten Vitalkapazität (FVC). Die European Respiratory Society hat Sollwertformeln für die IVC, in abhängig von Geschlecht, Alter und Körpergröße, aufgestellt (s. Quanjer [17, Tab. 6, S. 26], sowie Ulmer et al. [16, S. 76]).

Sollwertformel für Frauen in Liter: $\text{IVC} = 4,664h - 0,026a - 3,28$

Sollwertformel für Männer in Liter: $\text{IVC} = 6,103h - 0,028a - 4,654$

Körpergröße h in Meter, Alter a

Es wurden also zwei lineare Regressionen durchgeführt. Dabei wurde das relative Körpergewicht nicht als Kovariable mit in das Modell aufgenommen, obwohl starkes Übergewicht die spirometrischen Werte negativ beeinflusst. Die Formeln wurden aus Daten

von sowohl Rauchern als auch Nichtrauchern gewonnen. Trotz der Beeinträchtigung der Lungenfunktion durch Rauchen, wurde das Rauchverhalten ebenfalls nicht als Kovariable in das Modell mitaufgenommen. Mit Hilfe dieser Sollwerte lässt sich der gemessene Wert des IVC relativieren, indem man ihn durch seinen Sollwert teilt.

$$\text{IVC\%} = \frac{\text{gemessener IVC}}{\text{Sollwert-IVC}}$$

Da mit den absolut gemessenen Werten nicht erkennbar ist, ob ein Rückgang der Lungenfunktion durch eine Krankheit bedingt ist oder ob es sich um eine Alterserscheinung handelt, ist der IVC% dem IVC vorzuziehen. Moderne Spirometer geben diesen relativen Wert aus und verwenden dabei oft die obigen Sollwertformeln. Aufgrund einer großen Streuung können einzelne Werte des IVC% mehr als 130% oder weniger als 80% betragen, ohne das eine Krankheit vorhanden ist (s. Ulmer et al. [16, S. 73 u. S. 75-77]).

6.2.2 Einsekundenkapazität (FEV1)

Die *Einsekundenkapazität* ist das maximale Volumen in Liter, das in einer Sekunde durch eine kraftvolle Ausatmung ausgestoßen wird. Daher wird es auch als forciertes expiratorisches Volumen in einer Sekunde (FEV1) bezeichnet. Die European Respiratory Society hat auch für das FEV1 Sollwertformeln mittels linearer Regression berechnet (s. Quanjer [17, Tab. 6, S. 26], sowie Ulmer et al. [16, S. 76]).

Sollwertformel für Frauen in Liter: $\text{FEV1} = 3,95h - 0,025a - 2,6$
Sollwertformel für Männer in Liter: $\text{FEV1} = 4,301h - 0,029a - 2,492$
Körpergröße h in Meter, Alter a

Analog zu IVC% lässt sich FEV1% definieren. Die relativen Größen IVC% und FEV1% sind positiv korreliert und können sich dennoch erheblich unterscheiden, wie Ulmer et al. [16, S. 77] bemerken.

6.2.3 Tiffeneau-Index (FEV1%VC)

Der *Tiffeneau-Index* (FEV1%VC) ist der Prozentsatz des FEV1-Werts an der Vitalkapazität.

$$\text{FEV1\%VC} = \frac{\text{FEV1}}{\text{VC}}$$

Auch für diesen Wert geben Ulmer et al. [16, S. 77-78] altersabhängige Sollwerte an.

Alter	18-19	20-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64
FEV1%VC	82	80	78	77	75,5	74,5	73,5	72	70

Tabelle 5: Sollwerte des Tiffeneau-Index [16, Tab. 8.1, S.78]

6.2.4 Packungsjahr (PY)

Um das Rauchverhalten eines Menschen zu quantifizieren, verwendet man die Größe *Packungsjahr* (PY: *pack year*). Die Anzahl der Packungsjahre eines Rauchers ergibt sich aus dem Produkt der Raucherjahre und der Anzahl der täglich konsumierten Zigarettenpackungen. Dabei wird von 20 Zigaretten pro Packung ausgegangen. Ein Raucher, der also zehn Jahre lang eine halbe Packung täglich raucht kommt auf fünf Packungsjahre, genauso wie einer, der eine Packung pro Tag fünf Jahre lang raucht (s. Kroegel und Bonella [11, Tab. 2.1, S. 40]).

6.3 Interstitielle Lungenkrankheiten (ILD)

Bei den *interstitiellen Lungenerkrankungen* (ILD) handelt es sich um eine Gruppe von sowohl akuten als auch chronischen Krankheiten, die durch eine Entzündung sowie durch eine Fibrose, eine krankhafte Bindegewebevermehrung des für den Gasaustauschs zuständigen Lungenparenchyms, gekennzeichnet sind. Das Symptom der Atemnot unter Belastung ist ihnen gemeinsam. Diese Pneumonien (Lungenentzündungen) sind in ILDs mit bekannter Ursache und in ILDs mit unbekannter Ursache, die *idiopathischen interstitiellen Pneumonien* (IIP) zu unterscheiden. Die IIPs heißen idiopathisch, da ihre Pathogenese, d.h. die Entstehung und Entwicklung der Krankheit, unklar ist. Abbildung 1 gibt eine Übersicht über die Pneumonien, die im Folgenden näher erläutert werden. Um hinsichtlich der verschiedenen interstitiellen Lungenkrankheiten zu unterscheiden, gehört zu deren Diagnostik neben einer BAL-Auswertung auch ein radiologischer Befund durch Röntgen und ein histologischer Befund, also Ergebnisse aus einer Untersuchung des Lungengewebes im Rahmen einer Biopsie. Eine Lungenbiopsie kann unter Umständen aufgrund des damit verbundenen Risikos nicht immer durchgeführt werden (s. Kroegel und Bonella [11, S. 365]). Da sich diese Arbeit mit der statistischen Auswertung der BAL beschäftigt, wird auf Unterschiede bzgl. des histologischen und radiologischen Befunds nicht näher eingegangen. Im Fokus stehen vorallem epidemiologische Aspekte der ILDs und die Unterschiede in der BAL. Die verschiedenen idiopathischen interstitiellen Pneumonien wurden von der American Thoracic Society und der European Respiratory Society in einer gemeinsamen Veröffentlichung [2] klassifiziert (vgl. auch Kroegel und Bonella [11, S. 358]).

6.3.1 Idiopathische Lungenfibrose (IPF)

Die *idiopathische Lungenfibrose*, auch als idiopathische pulmonale Fibrose (IPF) bezeichnet, ist eine interstitielle Pneumonie unbekannter Ursache, die progressiv und chronisch verläuft. Da ihre Ursache nach Definition unbekannt ist, müssen für eine IPF-Diagnose andere IIPs sowie Umwelteinflüsse und Systemerkrankungen als Krankheitsverursacher ausgeschlossen werden. Die fortschreitende Fibrosierung führt durch die Vernarbung des Lungengewebes zu trockenem Husten und einer zunehmenden Luftnot bei Belastung. So sinkt die FVC im Mittel um 150-200 ml pro Jahr. Nur selten treten Allgemeinsymptome wie Gewichtsverlust und Müdigkeit auf. Bei der IPF handelt es sich um die am

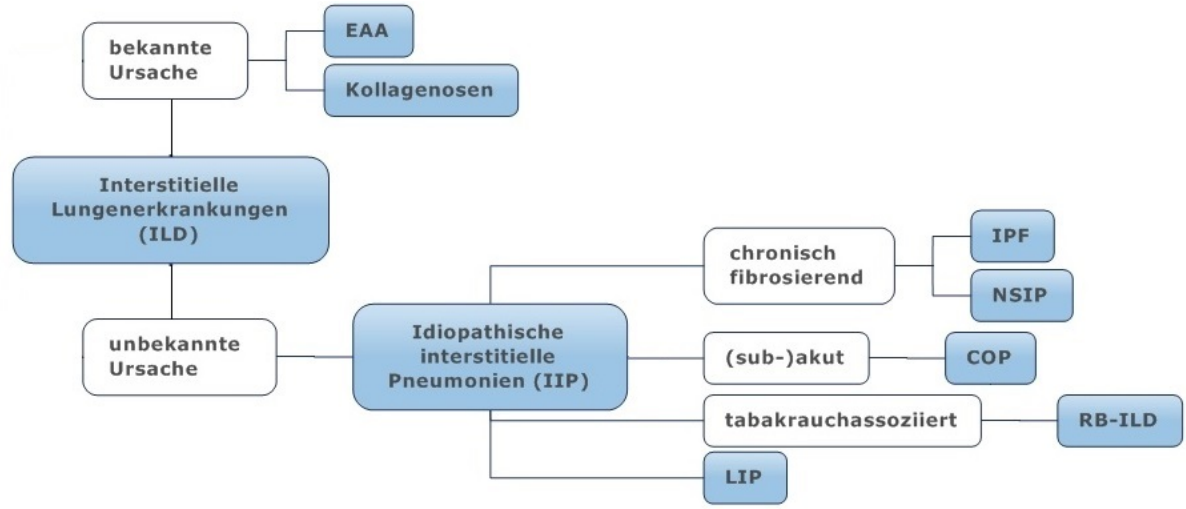

Abbildung 1: Interstitielle Lungenerkrankungen [11, Abb. 6.1, S. 358]

häufigsten auftretende IIP, die vorallem Männer und Raucher betrifft. Zudem weist sie mit einer durchschnittlichen Überlebenszeit von etwa 3-5 Jahren nach der Diagnosestellung gegenüber anderen IIPs eine höhere Mortalität auf. Der Umgang mit Metall- und Holzstaub gilt als Risikofaktor. Die Krankheit bricht vorallem zwischen dem 60. und 70. Lebensjahr aus und nur selten bei Patienten unter 50. Die BAL zeigt einen erhöhten Eosinophilen- und vorallem Neutrophilenanteil. Ist eine Lymphozytose festzustellen, so ist eher an eine EAA oder an eine NSIP zu denken (s. Kroegel und Bonella [11, S. 361-363]).

6.3.2 Nichtspezifische interstitielle Pneumonie (NSIP)

Aufgrund des histologischen Befunds wird die *nichtspezifische interstitielle Pneumonie* (NSIP) in zwei Subtypen unterschieden; die zelluläre und die fibrotische NSIP. Die BAL ist bei einer zellulären NSIP von Lymphozyten dominiert, bei einer fibrotischen von Neutrophilen und Eosinophilen. Die Symptome sowie der radiologische Befund ähneln denen der IPF. Es treten jedoch häufiger Allgemeinsymptome wie Müdigkeit, Gewichtsabnahme und eine leicht erhöhte Körpertemperatur auf. Die Krankheit bricht vorallem zwischen dem 40. und 50. Lebensjahr aus und betrifft Frauen etwas häufiger als Männer im Verhältnis 1.4:1. Raucher sind jedoch nicht häufiger betroffen. Die Mortalität ist etwas niedriger als bei der IPF, so beträgt die 10-Jahres-Überlebensrate im Mittel 70%. Eine sichere Diagnose kann nur mit Hilfe eines histologischen Befundes gestellt werden (s. Kroegel und Bonella [11, S. 365], sowie Müller-Quernheim [12, S. 92]).

6.3.3 Kryptogene organisierende Pneumonie (COP)

Das Krankheitsbild der *kryptogenen organisierenden Pneumonie* (COP), auch als Bronchiolitis obliterans mit organisierender Pneumonie (BOOP) bezeichnet, kann im Zu-

sammenhang mit Herztransplantationen, Autoimmunerkrankungen, Kollagenosen, etc. auftreten. In ihrer idiopathischen Form, d.h. wenn sie mit keiner dieser Gegebenheiten in Verbindung gebracht werden kann, tritt sie am häufigsten auf. Dabei sind Männer und Frauen gleich häufig davon betroffen, Raucher jedoch seltener als Nichtraucher. Als Symptome gelten Husten, Luftnot, Fieber, grippale Beschwerden, Gewichtsverlust und Nachtschweiß. Der Ausbruch der Krankheit findet vorallem zwischen dem 50. und 60. Lebensjahr statt. Die BAL zeigt häufig eine Lymphozytose und zusätzlich erhöhte Neutrophilen-, Eosinophilen- und Mastzellenwerte. Insgesamt ergibt sich ein ähnlicher BAL-Befund wie bei einer exogen allergischen Alveolitis (s. Kroegel und Bonella [11, S. 366], sowie Müller-Quernheim [12, S. 118]).

6.3.4 Lymphozytäre interstitielle Pneumonie (LIP)

Neben trockenem Husten und Atemnot sind vorallem lymphozytäre Infiltrate in der Lunge ein Indiz für die *lymphozytäre interstitielle Pneumonie* (LIP). Dabei dringen Lymphozyten in das Lungengewebe ein und lagern sich dort ab. Anhand der BAL lässt sich daher eine stark erhöhte Lymphozytenanzahl feststellen (s. American Thoracic Society und European Respiratory Society [2, S. 299]). Die LIP kann auch im Zusammenhang mit chronisch aktiver Hepatitis, einer Knochenmarktransplantation, dem HI-Virus, etc. auftreten. Bei HIV-negativen Menschen tritt sie selten vor dem 50. Lebensjahr auf (s. Müller-Quernheim [12, S. 125]).

6.3.5 Rispiratorische Bronchiolitis mit interstitieller Lungenerkrankung (RB-ILD)

Die *rispiratorische Bronchiolitis* mit interstitieller Lungenerkrankung (RB-ILD) ist eine interstitielle Lungenentzündung, begleitet von Atemnot und Husten, die ausschließlich bei (Ex-)Rauchern auftritt. Die RB-ILD manifestiert sich oft in der 4. Lebensdekade. Die 10-Jahres Überlebensrate beträgt über 70%. Die BAL ist durch Rauchermakrophagen in mehr als 80% der Fälle gelb-braun verfärbt. Sie zeigt eine hohe Gesamtzellenanzahl mit mehr als $25 \cdot 10^6$ Zellen pro Milliliter. Die Makrophagen dominieren mit über 80% den BAL-Befund (s. Kroegel und Bonella [11, S. 276-278], sowie Müller-Quernheim [12, S. 90]).

6.3.6 Exogen allergische Alveolitis (EAA)

Bei der *exogen allergischen Alveolitis* (EAA) kommt es aufgrund wiederholtem Einatmen von Allergenen (Antigene) zu einer allergischen Entzündungsreaktion des Lungenparenchyms. Taubenzüchter oder Wellensittichhalter können so durch die Inhalation von Allergenen, die im Zusammenhang mit Vögeln stehen, an EAA erkranken. Man spricht dann von einer Vogelhalterlunge. Des weiteren gibt es auch die Farmerlunge, verursacht durch eingeatmete Schimmelpilze von feuchtem Heu oder die Befeuchterlunge, die durch wiederholtes Inhalieren von Schimmelpilzen, Bakterien und Amöben, die sich zum Beispiel in Klimaanlagen ansiedeln, entsteht. Die EAA kann auch durch Medikamente ausgelöst werden. Sie kann akut mit grippeähnlichen Beschwerden zum Ausbruch kommen.

Reinigt ein Taubenzüchter mit einer Vogelhalterlunge seinen Taubenschlag, so kann er nach 4 bis 12 Stunden an Symptomen wie Fieber, Schüttelfrost, Gliederschmerzen und Husten leiden, die bald darauf wieder abklingen. In seiner chronischen Form, zum Beispiel wenn ein Wellensittichhalter kontinuierlich kleine Allergenmengen einatmet, sind die Symptome der EAA Atemnot unter Belastung, ein trockener Husten und ein chronisches Krankheitsgefühl begleitet von Appetitlosigkeit und Gewichtsverlust. Raucher sind seltener davon betroffen als Nichtraucher, während Männer häufiger davon betroffen sind als Frauen. Der Ausbruch der Krankheit findet vorallem zwischen dem 40. und 50. Lebensjahr statt. Anhand der BAL lässt sich eine deutlich erhöhte Gesamtzellen- sowie Lymphozytenanzahl erkennen. Des weiteren sind die Granulozyten und Mastozyten etwas erhöht. Liegt ein Normalbefund der BAL vor, so kann eine EAA so gut wie ausgeschlossen werden (s. Kroegel und Bonella [11, S. 367-370]).

6.3.7 Kollagenosen (KOL)

Bei *Kollagenosen* (Bindegewebserkrankungen) handelt es sich um verschiedene systemische Autoimmunerkrankungen. Als Systemerkrankungen können sie auf verschiedene Organe einen Einfluss haben und so auch eine ILD auslösen. Im Folgenden werden Kollagenosen als KOL abgekürzt. Im Gegensatz zu den anderen medizinischen Abkürzungen handelt es sich dabei jedoch nicht um eine etablierte, wie sie in der Standardliteratur benutzt wird. Die Symptome der interstitiellen Lungenerkrankung, die mit Kollagenosen assoziiert wird, entsprechen denen der EAA. Da die Patienten aufgrund der Kollagenose an rheumatischen Beschwerden leiden, kann eine Atemnot aufgrund der begrenzten Mobilität oft erst spät wahrgenommen werden (s. Müller-Quernheim [12, S. 106]).

7 Schildge-Datensatz: Statistische Modellierung

7.1 Ausgabewerte

Beschreibt die Variable G den Gesundheitszustand, den ein Patient mit Verdacht auf eine interstitielle Lungenerkrankung besitzt, so nimmt sie Werte in

$$\mathcal{G} = \{\text{IPF, COP, KOL, RBILD, EAA, LIP, NSIP, KON}\}$$

an, wobei mit KON die Kontrollgruppe der gesunden Individuen bezeichnet ist. In dieser Gruppe sind jene Patienten, die nach einer Untersuchung aufgrund des Verdachts auf Lungenentzündung für gesund befunden wurden. Es handelt sich also nicht um eine repräsentative Stichprobe der Gesamtbevölkerung. Im Sinne der Bayesschen Entscheidungstheorie werden daher nicht die A-posteriori-Wahrscheinlichkeiten, dass ein beliebiges Individuum aus der Bevölkerung an einer bestimmten ILD leidet, modelliert. Sie sind danach bedingt, dass sich der Patient in einer pneumologischen Facharztpraxis untersuchen lässt. Insgesamt wurden 894 Patienten untersucht, davon 183 mit IPF, 191 mit COP, 147 mit KOL, 97 mit RBILD, 118 mit EAA, 41 mit LIP und 29 mit NSIP. Zur Kontrollgruppe gehören 88 Patienten.

7.2 Eingabewerte und Normalverteilung

Folgende erfasste Größen sind mögliche Eingabewerte.

1. Epidemiologie:
   a) Alter X_A,
   b) Geschlecht X_{mw},
   c) Rauchverhalten X_{PY} in Packungsjahre.
2. BAL-Parameter:
   a) Gesamteiweiß, bzw. Protein X_P und Albumin X_{Alb} in der BAL in mg/l und die transformierten Daten $X_{lP} = \log X_P$ und $X_{lAlb} = \log X_{Alb}$. Albumin ist ein bestimmtes Protein, das im Gesamteiweiß mitberücksichtigt wird.
   b) Zellendichte X_Z in Zellen pro μl, $X_{lZ} = \log X_Z$,
   c) Makrophagenanteil X_M in Prozent, $X_{lM} = \log X_M$,
   d) Lymphozytenanteil X_L in Prozent, $X_{lL} = \log(X_L + 1)$. Dabei muss für eine Logarithmierung der Lymphozyten eine Eins hinzu addiert werden, da sie insbesondere in der Kontrollgruppe oft nicht beobachtetet werden.
   e) Granulozytenanteil X_G (Neutrophile, Eosinophile und Mastzellen) in Prozent, $X_{lG} = \log(X_G + 1)$.

3. Spirometrie:
   a) Relative inspiratorische Vitalkapazität $X_{IVC\%}$,
   b) Relative Einsekundenkapazität $X_{FEV1\%}$,
   c) Tiffeneau-Index X_{Tiff}.

Im Anhang befinden sich Boxplots und andere Visualisierungen (s. Abbildung 8-12) der Daten. Mit Hinblick auf die lineare Diskriminanzanaylse, die die Normalverteilung der Daten annimmt, gibt Tabelle 6 einen Überblick über die p-Werte, die sich aus einem Shapiro-Wilk-Test auf Normalverteilung ergeben. Tabelle 7 gibt die p-Werte des Royston-Tests auf multivariate Normalverteilung an.

$X\|G=k$	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
X_A	1.8e-4	0.0022	3e-4	0.27	0.0087	0.0021	0.0049	0.0064
X_{PY}	6.9e-16	8e-18	4.6e-18	1.1e-4	6.9e-18	8.9e-9	5.1e-7	9.7e-14
X_{lP}	0.16	0.0033	5.4e-4	0.0027	0.034	0.5	0.075	0.25
X_{lAlb}	0.74	0.18	3.7e-4	0.0011	0.78	0.55	0.16	0.89
X_{lZ}	2e-09	4.3e-06	7.1e-09	2.3e-4	6.5e-4	0.22	0.039	8.1e-9
X_{lM}	3.9e-19	4.3e-11	9.6e-12	6.8e-19	3.4e-4	1.2e-4	4.8e-5	9.8e-8
X_M	8.1e-12	1.9e-4	7e-5	3.4e-14	2.4e-4	0.67	0.15	6.8e-6
X_{lL}	0.44	8.4e-9	6.1e-4	0.0029	5.7e-10	0.25	0.83	0.017
X_{lG}	0.28	0.11	0.048	4.4e-4	0.21	0.09	0.31	9.6e-4
$X_{IVC\%}$	0.24	0.91	0.49	0.23	0.92	0.18	0.24	0.89
$X_{FEV1\%}$	0.83	0.32	0.15	0.34	0.9	0.11	0.52	0.025
X_{Tiff}	4.6e-15	5.7e-4	0.13	0.11	7e-7	5.9e-8	0.99	0.018

Tabelle 6: p-Werte des Shapiro-Wilk-Tests auf univariate Normalverteilung

$X\|G=k$	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
X_{lL}, X_{lG}	0.41	2.1e-8	4e-4	1.3e-5	2.1e-9	0.12	0.58	3.4e-4
X_{lAlb}, X_{lL}, X_{lG}, $X_{IVC\%}$	0.43	4.8e-6	4.4e-4	5.3e-7	1.3e-6	0.49	0.33	0.0057

Tabelle 7: p-Werte des Royston-Tests auf multivariate Normalverteilung (Auswahl)

Unter den Eingabewerten gibt es zahlreiche Abhängigkeiten. Tabelle 30 im Anhang gibt durch die empirischen Pearson-Korrelationskoeffizienten einen Eindruck davon. Dabei wurden die Koeffizienten mit den Daten aller Gruppen berechnet. Es lässt sich nicht ausschließen, dass in einigen Gruppen die entsprechenden Eingabewerte stärker miteinander korreliert sind als in anderen. Protein und Albumin sind mit einer Korrelation von $\hat{\rho}(X_P, X_{Alb}) = 0.91$ linear abhängig, da Albumin zu dem Gesamtprotein dazuzählt. Ein Ziel dieser Arbeit ist es, statistisch zu begründen, ob sich der Protein- bzw. Albumingehalt in der BAL als Eingabewert für Klassifikationsverfahren eignet. Albumin und Protein codieren dabei ähnliche Informationen, wie Abbildung 9 zeigt. Es stellt sich die

Frage, welcher der beiden sich besser eignet. Eine erste Aussage lässt sich durch Tabelle 6 treffen; die Normalverteilungsannahme der logarithmierten Albuminwerte wird weniger oft verworfen als die des Proteins.
Je stärker die Entzündung der Lunge vorangeschritten ist, desto höher sind die Werte von Zelldichte, Protein, Albumin, Lymphozyten und Granulozyten. Diese Entzündungswerte sind daher nach Tabelle 30 deutlich positiv korreliert. Ist durch eine Entzündungsreaktion der Lymphozyten- und Granulozytenanteil in der BAL erhöht, so ist der Makrophagenanteil entsprechend niedrig. Daher ist Letzterer mit den Entzündungswerten stark negativ korreliert.
Die Fähigkeit der Lunge, so viel Luft wie möglich durch langsames Einatmen aufnehmen zu können (inspiratorische Vitalkapazität), korreliert stark mit der maximalen Luft, die sie in einer Sekunde so schnell wie möglich ausatmen kann (expiratorisches Volumen in einer Sekunde). Diese Aussage ist in Form des Korrelationskoeffizienten $\hat{\rho}(X_{IVC\%}, X_{FEV1\%}) = 0.87$ quantifiziert.

7.3 Zytozentrifuge und Beta-Binomialverteilung

Das Differentialbild der BAL wird mit Hilfe einer Zytozentrifuge erstellt. Diese Datenerhebung wird in der Richtlinie der American Thoracic Society beschrieben [14, S. 1005-1014]. Darin wird diskutiert, dass zumindest bei gesunden Menschen die Immunzellenanteile in der BAL asymmetrisch und damit nicht normalverteilt sind. In diesem Kapitel wird mit der Beta-Binomialverteilung eine Parametrisierung zur statistischen Modellierung des BAL-Differentialbilds vorgeschlagen. Dabei wurden für den Datensatz jeweils 600 Zellen in der BAL mikroskopisch untersucht und in Makrophagen, Lymphozyten, Granulozyten (Eosinophile, Neutrophile, Mastozyten), Plasmazellen und Monozyten differenziert. Der Plasmazellen- und Monozytenanteil X_{PM} ist grundsätzlich sehr niedrig. Beide besitzen keinerlei diagnostische Relevanz und wurden daher nicht im Datensatz gespeichert. Die anderen relativen Anteile liegen auf Tausendstel gerundet vor. Aufgrund ihrer Erhebung gilt folgender systematischer Zusammenhang:

$$X_M + X_L + X_G + X_{PM} = 100\%. \tag{7.1}$$

Dominieren die Entzündungswerte (X_L, X_G) die BAL, so liegen die entsprechenden Datenpunkte in der Nähe des Randes der Menge $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 = 100\}$. Die transformierten Daten erfüllen folgenden nichtlinearen systematischen Zusammenhang

$$X_M = 103 - \exp(X_{lL}) - \exp(X_{lG}) - \exp(\log(X_{PM} + 1)),$$

wie in Abbildung 7 illustriert. Eine naheliegende statistische Modellierung der BAL-Differentialzytologie ist es, die Anzahl der Lymphozyten an den insgesamt 600 ausgezählten Zellen durch eine Binomialverteilung zu beschreiben. Ein Modellierungsproblem ist dabei, dass die Daten in Prozent auf eine Nachkommastelle gerundet vorliegen. Durch die Transformation $10X_L$ werden diese Zahlen wieder in natürliche überführt. Ein Blick

auf Abbildung 10 zeigt, dass der Lymphozytenanteil asymmetrisch verteilt ist und in einigen Gruppen eine stärkere Streuung aufweist. Dieses Phänomen lässt sich stochastisch abbilden, indem man es der Erfolgswahrscheinlichkeit der Binomialverteilung erlaubt, selbst eine Zufallsvariable zu sein. Man spricht in diesem Fall von einer Mischverteilung. Medizinisch lässt sich diese Modellierung dadurch begründen, dass sich die Patienten in unterschiedlichen Behandlungs- und Krankheitsstufen befinden. Für eine Beta-verteilte Erfolgswahrscheinlichkeit $P \sim \text{Beta}(\alpha, \beta)$ ergibt sich so für einen Patienten mit $P = p$ die Verteilung des transformierten Lymphozytenanteils zu $10X_L|P = p \sim \text{Bin}(1000, p)$. Diese Konstruktion führt zur Definition der Beta-Binomialverteilung (s. Johnson, Kemp und Kotz [10, S. 253-256]).

Definition 7.1. Eine Variable X heißt *Beta-binomialverteilt*, i. e. $X \sim \text{BeB}(n, \alpha, \beta)$, falls gilt

$$\mathbb{P}(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} dp, \; x = 0, \ldots, n \text{ und } n \in \mathbb{N},$$

wobei mit $B(\alpha, \beta)$ die Betafunktion für $\alpha, \beta > 0$ bezeichnet ist. Der Erwartungswert und die Varianz der Beta-Binomialverteilung ist gegeben durch

$$\mathbb{E}X = \frac{n\alpha}{\alpha + \beta} \text{ und } \mathbb{V}X = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Die Histogramme in Abbildung 2 und 3 veranschaulichen die Verteilungen der transformierten Lymphozytendaten innerhalb der verschiedenen Patientengruppen. Die Klassenbreite entspricht bei allen Darstellungen 20. In rot ist die Dichte der geschätzten, diskreten Beta-Binomialverteilung angedeutet, deren Parameter mit Hilfe des R-Pakets *VGAM* durch eine Maximum-Likelihood-Schätzung berechnet wurden. Analog dazu sind im Anhang durch Abbildung 13 und 14 die Granulozytendaten visualisiert. Tabelle 8 gibt die geschätzten Parameter an mit denen sich die bedingten Erwartungswerte und Varianzen von X_L und X_G schätzen lassen durch

$$\hat{\mu} = \frac{n\hat{\alpha}}{10(\hat{\alpha} + \hat{\beta})} \text{ und } \hat{\sigma}^2 = \frac{n\hat{\alpha}\hat{\beta}(\hat{\alpha} + \hat{\beta} + n)}{100(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)} \text{ mit } n = 1000.$$

Insgesamt stimmen die Resultate mit den wissenschaftlichen Erkenntnissen über interstitielle Lungenerkrankungen, wie sie in Kapitel 6.3 wiedergegeben werden, überein. Die Krankheiten COP und EAA zeichnen sich durch einen stark erhöhten Lymphozytenanteil in der BAL aus, während dieser Entzündungswert für IPF-Patienten nur leicht erhöht ist. Dabei sind diese und folgende Aussagen im Sinne eines Erwartungswertes zu verstehen. Bei an RBILD erkrankten (Ex-)Rauchern ist der Lymphozytenanteil im Allgemeinen sogar etwas geringer als in der Kontrollgruppe, da Rauchermakrophagen das Bild der BAL dominieren. Bei der Schätzung der Verteilung der Granulozyten in IPF und COP ist ein Unterschied zwischen der geschätzten Wahrscheinlichkeit und der tatsächlichen relativen Wahrscheinlichkeit zur ersten Klasse dazuzugehören, zu beobachten. Gleichzeitig sind stark erhöhte Werte sichtbar (s. Abbildung 13 links und rechts oben). Es liegt nahe, dass es sich dabei um extreme Beobachtungen handelt, die als Ausreißer die Schätzung der Verteilungsparameter verfälschen.

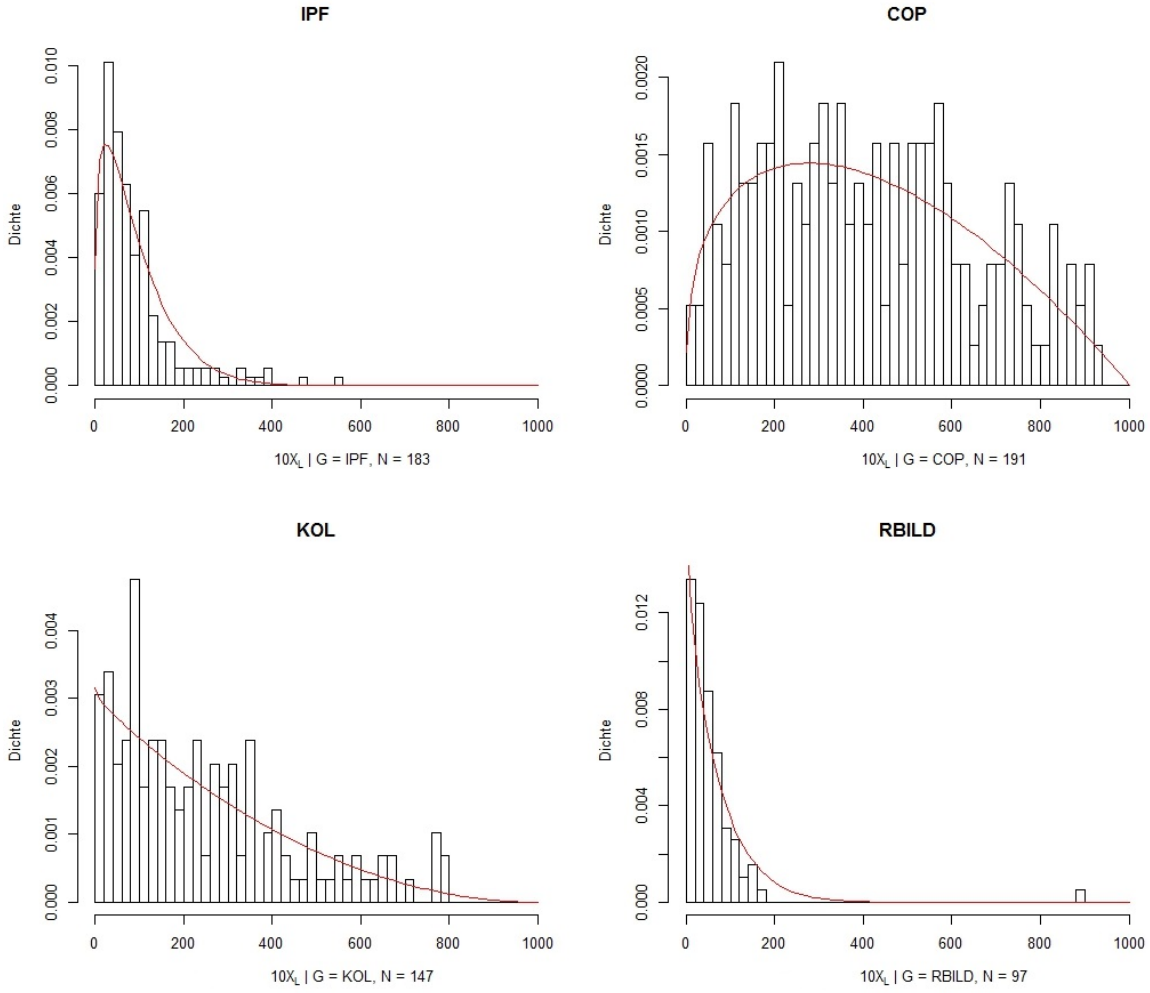

Abbildung 2: Beta-Binomialverteilung der Lymphozyten I

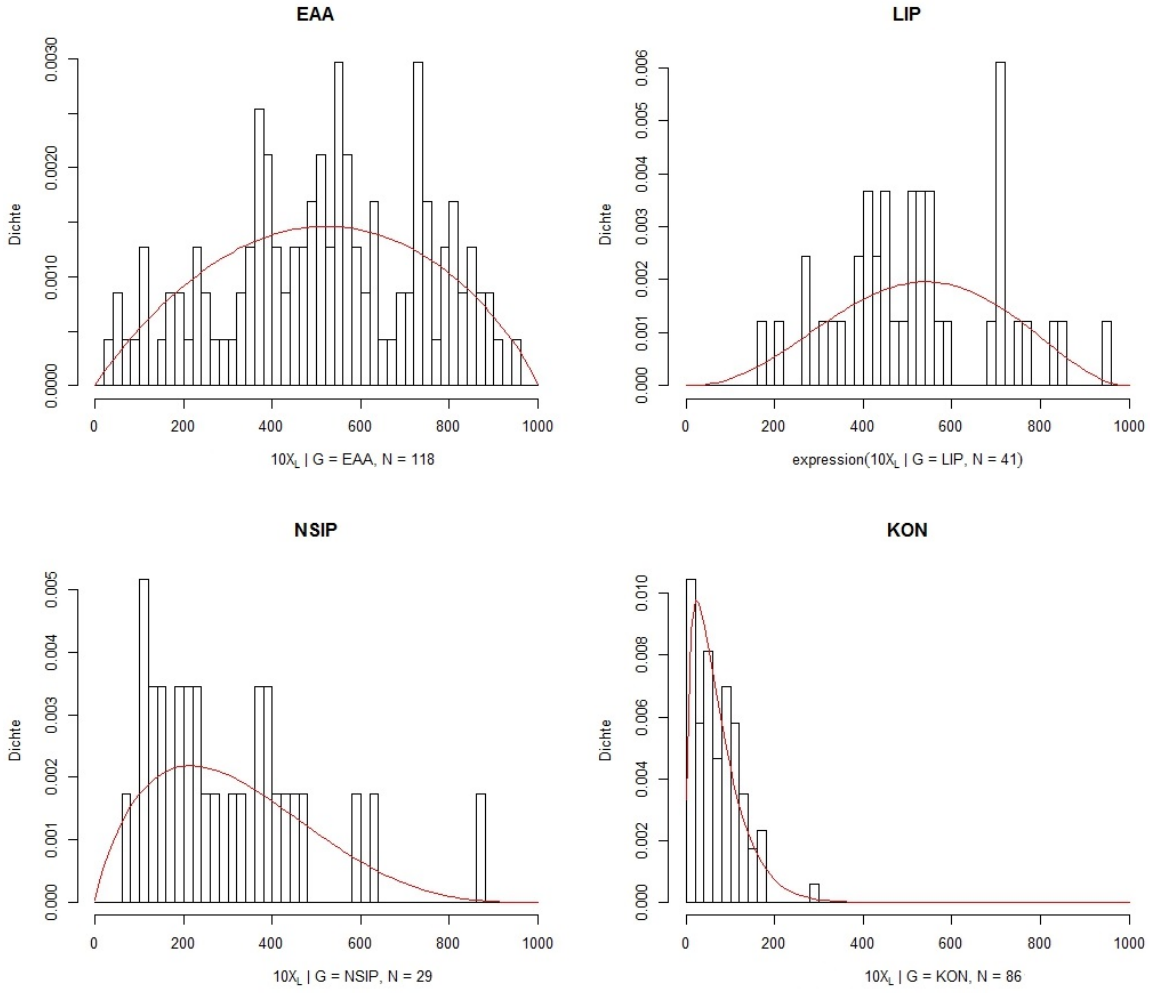

Abbildung 3: Beta-Binomialverteilung der Lymphozyten II

	$X_L \mid G = k$				$X_G \mid G = k$			
Gruppe k	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\alpha}$	$\hat{\beta}$
IPF	9.3	57.3	1.28	12.58	18.0	228.3	0.99	4.51
COP	41.1	559.4	1.37	1.97	15.6	207.0	0.84	4.55
KOL	25.0	379.8	0.99	2.97	10.3	98.8	0.87	7.59
RBILD	6.8	42.7	0.95	13.08	4.9	27.1	0.81	15.68
EAA	51.1	519.2	1.96	1.87	11.9	115.2	0.97	7.2
LIP	52.9	333.5	3.45	3.07	9.4	93.3	0.77	7.41
NSIP	31.2	309.0	1.87	4.12	14.2	138.3	1.11	6.74
KON	7.1	31.7	1.43	18.79	2.1	3.8	1.17	54.78

Tabelle 8: Parameterschätzung für Lymphozyten und Granulozyten

7.4 Rauchen, Geschlecht und Binomialverteilung

$X \mid G = k, B = 1$	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
X_{PY}	3.1e-6	3.4e-5	0.37	1.1e-4	6.6e-4	0.27	0.15	7.4e-4
$\log X_{PY}$	1.4e-4	0.015	0.001	2.9e-4	0.22	0.61	0.17	0.011
(Ex-)Raucher (%)	44.8	38.2	26.5	100	24.6	22	34.5	34.1
Nichtraucher (%)	50.8	56	70.7	0	75.4	65.9	62.1	65.9
Verhältnis	0.9:1	0.7:1	0.4:1	-	0.3:1	0.3:1	0.6:1	0.5:1

Tabelle 9: Rauchen

	(Ex-)Raucher	Nichtraucher	Σ
Männer	274	200	474
Frauen	95	296	391
Σ	369	496	865

Tabelle 10: Rauchen und Geschlecht

Ob ein untersuchtes Individuum ein (Ex-)Raucher ist oder nicht, lässt sich durch eine binomialverteilte Zufallsvariable $B \sim \text{Bin}(1, p)$ beschreiben, wobei Raucher zu sein, bzw. gewesen zu sein als Erfolg codiert ist. Rauchen und Nichtrauchen gelten für interstitielle Lungenkrankheiten epidemiologisch betrachtet als Risikofaktoren, wie in Kapitel 6.3 beschrieben. Somit ist die Zufallsvariable B stochastisch abhängig von der Gruppe. Die p-Werte des Shapiro-Wilk-Tests auf Normalverteilung der (transformierten) Packungsjahre der Raucher sind in Tabelle 9 aufgeführt. Des Weiteren sind die Anteile an Nichtrauchern und Rauchern in den jeweiligen Gruppen angegeben, wobei bei 29 Patienten keine Daten zum Rauchverhalten erhoben wurden. Dadurch summieren sich die Werte nicht zu hundert Prozent auf und der Raucherprozentsatz ist nicht die korrekte Schätzung der Erfolgswahrscheinlichkeit der Binomialverteilungen. Der Männer- und Raucheranteil in der Gruppe der an IPF Leidenden ist, wie in der Literatur beschrieben, erhöht. Die Hypothese der Unabhängigkeit von Geschlecht und Rauchersein kann nach Tabelle 10 durch einen Chi-Quadrat-Test signifikant verworfen werden. So beträgt der Männeranteil in RBILD 73.2%. Aus dem

Datensatz ergibt sich dasselbe Verhältnis von Männer zu Frauen in der NSIP-Gruppe wie es in der Literatur angegeben ist (1.4:1). Bei der späteren statistischen Klassifikation der Daten sind solche Abhängigkeiten insofern wichtig, als dass man hinterfragen muss, ob ein gewählter Eingabewert kausal mit der Krankheit zusammenhängt oder ob er nur mit einem anderen korreliert, der dies tut. Sind also wirklich Männer die Risikogruppe oder vielmehr die Raucher?

Diagnose	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
männlich (%)	69.4	60.7	28.6	73.2	51.7	58.5	58.6	42
weiblich (%)	30.6	39.3	71.4	26.8	48.3	41.5	41.4	58
Verhältnis	2.3:1	1.5:1	0.4:1	2.7:1	1.1:1	1.4:1	1.4:1	0.7:1

Tabelle 11: Geschlecht

8 Statistische Klassifikation in interstitielle Lungenkrankheiten

8.1 Relevanz der Eingabewerte

In diesem Kapitel werden die Ergebnisse von unterschiedlichen an den Datensatz angepassten Klassifikationsregeln ausgewertet. Durch Vorwärtsselektion der Eingabewerte wurden in Tabelle 12 die Variablen schrittweise gewählt, die den .632-Schätzer minimieren. Zur Berechnung aller Größen wurde die 0/1-Verlustfunktion verwendet. Dabei wurde jeweils die Klassifikationsmethode der linearen Diskriminanzanalyse angewendet. In einem ersten Schritt wurde der logarithmierte Lymphozytenanteil als die relevanteste Größe identifiziert, um die acht Gruppen voneinander zu unterscheiden. Die Daten lassen sich so grob in die Gruppen, in denen dieser Entzündungswert kaum (IPF, RBILD, KON), stark (COP, EAA, LIP) und moderat erhöht ist (KOL, NSIP), unterteilen (s. Abbildung 2, 3 und 11). Diese Aussage, so wie auch folgende, sind im Sinne eines Erwartungswertes bzw. Medians zu verstehen. Im Allgemeinen ist z. B. nicht auszuschließen, dass ein EAA-Patient einen BAL-Normalbefund aufweist. Grundsätzlich sind so hohe, zweistellige Trainingsfehler zu erwarten. Festzuhalten bleibt, dass die drei weiteren Bestplaziertesten, bzgl. der Minimierung des .632-Schätzers; Protein-, Albumingehalt und Makrophagenanteil, jene Variablen sind, die mit dem Lymphozytenanteil am stärksten korreliert sind. Verwendet man für eine Modellanpassung die Lymphozyten und zusätzlich jeweils eine dieser drei Eingabewerte, so erhält man kaum einen Informationsgewinn. Die drei jeweiligen Modelle sind unter Schritt 2 die drei Schlechtplatziertesten. Daraus erschließt sich, dass die vier Variablen Lymphozyten, Protein, Albumin und Makrophagen eine ähnliche Information codieren. Sie können zwischen den Gruppen mit kaum (IPF, RBILD, KON), stark (COP, EAA, LIP) und moderat erhöhten (KOL, NSIP) Entzündungswerten unterscheiden (vgl. Abbildung 9, 10 und 11). Mit Hilfe des Granulozytenanteils wie er im 2. Schritt gewählt wird, lässt sich z. B. IPF von der Kontrollgruppe gut trennen. Des weiteren ist der Median des Granulozytenanteils in der Gruppe LIP etwas geringer als in EAA und COP, da hier die Lymphozyten die BAL stärker dominieren. Durch die zusätzliche Information von Packungsjahren und Alter lassen sich zusätzlich vorallem RBILD-Patienten identifizieren, da sie starke Raucher und relativ jung sind. Die Aufnahme von der prozentualen inspiratorischen Vitalkapazität und Proteingehalt führt zu weiteren marginalen Verbesserungen des .632-Schätzers bis er sich schließlich nur durch weitere aufgenommene Eingabewerte erhöhen würde. Zusätzlich wurde die Leave-One-Out-Kreuzvalidierung (CV) sowie die 5-fache Kreuzvalidierung (5-CV) berechnet.

Abbildung 4 beschreibt anschaulich die Trennung des Datensatzes nach den Eingabewerten, gemäß eines mit dem Gini-Index konstruierten Klassifikationbaums. Der Baum wurde dabei mit dem geschätzten Komplexitätsparameter $\hat{\alpha} = 0.008535$ nach Kapitel 4.4.3 reduziert. Die Zahlen in den Knoten geben jeweils an, wie viele Daten zu den Gruppen gehören. Oben v. l. n. r.: IPF, COP, KOL und RBILD. Unten: EAA, LIP, NSIP

Eingabewerte	N	err	CV	5-CV	$\widehat{\text{Err}}^{.632}$
1. Schritt					
X_A	894	0.729	0.729	0.739	0.737
X_{mw}	894	0.74	0.74	0.757	0.747
X_{PY}	865	0.749	0.761	0.762	0.763
X_{lP}	894	0.682	0.682	0.683	0.686
X_{lAlb}	886	0.685	0.685	0.681	0.688
X_{lZ}	889	0.746	0.746	0.745	0.745
X_{lM}	892	0.697	0.698	0.695	0.697
X_{lL}	892	**0.643**	**0.652**	**0.649**	**0.649**
X_{lG}	893	0.765	0.773	0.763	0.761
$X_{IVC\%}$	824	0.729	0.729	0.745	0.733
$X_{FEV1\%}$	832	0.767	0.77	0.778	0.774
X_{Tiff}	802	0.737	0.742	0.747	0.743
2. Schritt					
X_{lL}, X_A	892	0.587	**0.594**	0.602	0.592
X_{lL}, X_{mw}	892	0.631	0.639	0.65	0.635
X_{lL}, X_{PY}	863	0.612	0.62	0.626	0.619
X_{lL}, X_{lP}	892	0.632	0.64	0.636	0.639
X_{lL}, X_{lAlb}	884	0.638	0.645	0.649	0.642
X_{lL}, X_{lZ}	887	0.623	0.643	0.639	0.635
X_{lL}, X_{lM}	892	0.633	0.638	0.649	0.643
X_{lL}, X_{lG}	892	**0.575**	0.596	**0.582**	**0.584**
$X_{lL}, X_{IVC\%}$	822	0.594	0.609	0.609	0.604
$X_{lL}, X_{FEV1\%}$	830	0.623	0.637	0.633	0.63
X_{lL}, X_{Tiff}	800	0.609	0.622	0.616	0.622
3. Schritt					
X_{lL}, X_{lG}, X_{PY}	863	**0.526**	**0.534**	**0.528**	**0.54**
4. Schritt					
$X_{lL}, X_{lG}, X_{PY}, X_A$	863	**0.505**	**0.511**	**0.516**	**0.514**
5. Schritt					
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}$	797	**0.479**	**0.491**	0.499	**0.497**
6.Schritt					
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}$	797	**0.46**	**0.478**	**0.488**	**0.485**
Stopp					
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{mw}$	797	0.476	0.492	0.497	0.491
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{lAlb}$	793	0.469	0.491	0.493	0.491
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{lZ}$	794	0.475	0.487	0.475	0.49
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{lM}$	797	0.471	0.486	0.48	0.493
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{FEV1\%}$	789	0.464	0.484	0.487	0.487
$X_{lL}, X_{lG}, X_{PY}, X_A, X_{IVC\%}, X_{lP}, X_{Tiff}$	773	0.464	0.484	0.505	0.487

Tabelle 12: LDA

und KON. Die Klassifikationsregel entscheidet sich für die am häufigsten vorkommende Gruppe in einem Knoten; die Knotenüberschrift. Die Wichtigkeit der Variablen Lymphozyten, Granulozyten und Packungsjahre wird auch hier bestätigt. Mit Hilfe eines niedrigen Lymphozytenanteils können die Gruppen IPF, RBILD und KON identifiziert werden, die sich wiederum durch die Granulozyten und durch die Packungsjahre gut trennen lassen. Der Trainingsfehler entspricht $484/894 = 0.541$. Vorallem im rechten Blatt entstehen 285 Fehlklassifikationen, indem die Gruppen COP, KOL, EAA, LIP und NSIP nicht mehr weiter voneinander getrennt werden und so z. B. alle Beobachtungen die zu LIP zugehörig sind, in die Gruppe COP eingeteilt werden. Der Konstruktionsalgorithmus des Klassifikationsbaums stellt hier die Vermeidung einer Erhöhung der Modellkomplexität über die Senkung des Trainingsfehlers. Verwendet man die Kreuzentropie zur Bestimmung eines Baumes ergibt sich ein analoges Bild, wobei der Datensatz durch die Lymphozyten bei 2.9 anstatt 2.8, durch die Granulozyten bei 1.8 anstatt 1.6 und durch die Packungsjahre bei gleichbleibenden 2.5 PY getrennt wird.

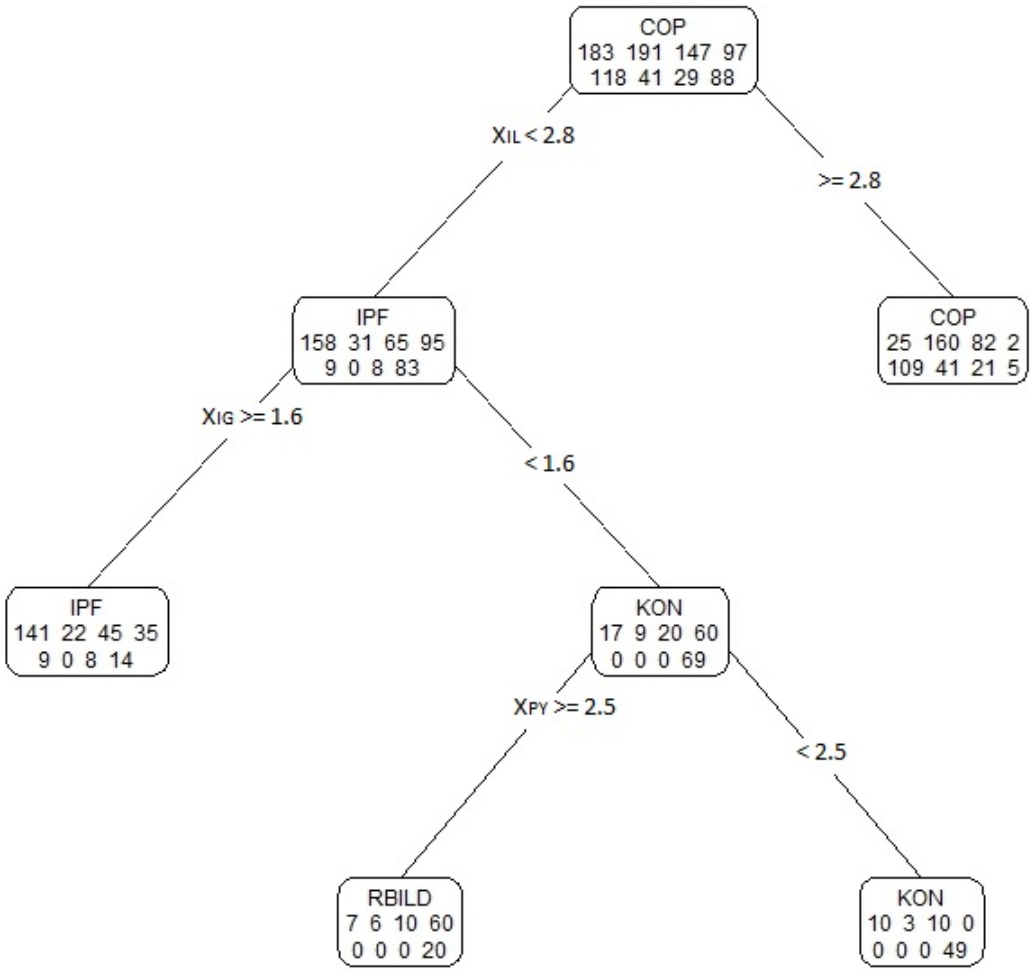

Abbildung 4: Klassifikationsbaum (Gini-Index)

	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
IPF	**140**	23	38	16	8	0	8	6
COP	16	**142**	47	1	79	36	14	2
KOL	5	13	**39**	0	9	3	7	3
RBILD	11	4	9	**74**	0	0	0	19
EAA	0	3	2	0	**22**	2	0	0
LIP	0	0	0	0	0	**0**	0	0
NSIP	0	0	0	0	0	0	**0**	0
KON	11	6	12	6	0	0	0	**58**
Summe	183	191	147	97	118	41	29	88
Fehlerrate	23.5%	25.7%	73.5%	23.7%	81.4%	100%	100%	34.1%

Tabelle 13: Klassifikationstabelle (Bagging)

Durch die Bagging-Methode unter Verwendung von $B = 50$ Bootstrap-Datensätzen und derselben Anzahl an Klassifikationbäumen fällt der Trainingsfehler auf $419/894 = 0.469$. Zur Konstruktion wurde der Gini-Index verwendet. Eine Übersicht der Ergebnisse, wie das Komitee dieser Bäume die Daten einteilt, wird in Tabelle 13 gegeben. Alle Datensätze gehören zur Gruppe ihrer Spaltenbezeichnung und wurden durch den Bagging-Klassifikator in die Gruppen der Zeilen eingeteilt. Die Diagonale gibt folglich die Zahlen der korrekt klassifizierten Patienten an. Auffällig sind die hohen Fehlerraten in KOL, EAA, LIP und NSIP. Die Gruppen LIP und NSIP werden dabei aufgrund geringer A-priori-Wahrscheinlichkeiten von den häufigeren Krankheiten mit hohen Entzündungswerten marginalisiert. Die Klassifikationsbäume bevorzugen es, diese seltenen Gruppen vollständig fehlzuklassifizieren. Eine Verdoppelung von B reduziert den Trainingsfehler mit dieser Methode lediglich auf 0.453. Mit Hilfe des R-Pakets *adabag* wurde die relative Relevanz der Variablen, bzgl. des Informationsgewinns nach dem Gini-Index, wie von Alfaro, Gamez und Garcia [1, S. 2] programmiert, in Tabelle 14 berechnet. Als die wichtigsten vier Variablen werden auch hier Lymphozyten, Granulozyten, Packungsjahre und Alter erkannt.

X_{lL}	X_{lG}	X_{PY}	X_A	X_{lP}	X_{lM}	$X_{IVC\%}$	X_{lAlb}	X_{Tiff}	X_{mw}	$X_{FEV1\%}$	X_{lZ}
36	20.5	16.3	8.3	3.8	3.3	2.8	2.7	2.4	1.5	1.3	1.1

Tabelle 14: Relevanz bzgl. des Informationsgewinns in % (Bagging)

8.2 Vergleich der Methoden zur Klassifizierung in IPF, KOL und EAA

Kapitel 8.1 zeigt die Grenzen des Möglichen bei der Kategorisierung der Daten in die acht Gruppen. Seltene Krankheiten wie LIP und NSIP werden durch ihre geringe A-priori-Wahrscheinlichkeiten kaum erkannt. Des weiteren besitzen die häufigen, von hohen Entzündungswerten geprägten Krankheiten EAA und COP einen ähnlichen BAL-Befund und sind ohne radiologischen und histologischen Befund schwer voneinander zu unterscheiden. Diese Gegebenheiten führen zu strukturell hohen Klassifikationsfehlern in Kapitel 8.1. Aufgrund einer hohen Mortalität ist es ein medizinisch folgenschwerer Fehler, einen IPF-Patienten falsch zu diagnostizieren. Die Beweislage ist jedoch durch BAL, Röntgenbild und Biopsie oft nicht eindeutig. In der Praxis kommt es dabei oft zu Zweifeln bei der Abgrenzung zwischen IPF, KOL, EAA und NSIP. Aufgrund des relativ niedrigen Stichprobenumfangs der Gruppe NSIP werden in diesem Kapitel die Klassifikationsverfahren auf die drei übrigen Gruppen angewendet und die Ergebnisse miteinander verglichen.
Tabelle 15 zeigt, dass auch für diese Gruppen im Rahmen einer linearen Diskriminanzanalyse der Lymphozytenanteil in der BAL als erstes gewählt wird. Auffällig ist hier, dass in einem zweiten Schritt die Hinzunahme von spirometrischen Daten das Modell weiter verbessern würde, die Daten jedoch nicht von allen Patienten erhoben wurden. Hier ist fraglich, ob diese Eingabewerte wirklich relevant sind oder ob die Fehler nur durch eine Reduktion des Datensatzes sinken. Daher werden sie im Folgenden bei der Variablenselektion nicht mitberücksichtigt. Die Wahl der relativen inspiratorischen Vitalkapazität in Tabelle 12 als Eingabewert ist insofern ebenfalls fraglich. Andere komplexere Modelle führen zwar zu einer geringfügigen Verbesserung des .632-Schätzers, erhöhen aber die Kreuzvalidierungen und insbesondere die Trainingsfehler, weswegen hier das Modell, das nur die Lymphozyten als Eingabewert besitzt, zu empfehlen ist.
Bei einer quadratischen Diskriminanzanalyse ist das Bild etwas klarer (s. Tabelle 16). Hier werden die Variablen Lymphozyten, Granulozyten, Geschlecht und Alter selektiert. Die Packungsjahre sind im Gegensatz zu Kapitel 8.1 nicht mehr von Interesse, da die Raucherkrankheit RBILD nicht mehr erkannt werden muss. Es besteht ein systematischer Unterschied zwischen den Geschlechtern in diesen drei Gruppen. In IPF sind vorallem Männer, in KOL vorallem Frauen, wohingegen sich beide Geschlechter in EAA die Waage halten (s. Tabelle 11).
Dementsprechend verwenden die Klassifikationsbäume, konstruiert mit dem Gini-Index (s. Abbildung 5) und mit der Kreuzentropie (s. Abbildung 6), das Geschlecht zur Unterscheidung von IPF und KOL mit einem Trainingsfehler von 0.272 bzw. 0.288. Zu bemerken ist, dass im rechten Teilbaum von Abbildung 5, die Lymphozyten zur weiteren Trennung von EAA und KOL verwendet werden, während diese Aufgabe in Abbildung 6 der Proteingehalt übernimmt. Ein weiteres Indiz, das untermauert, dass Protein und Lymphozyten dieselbe Information codieren.
Bagging und Boosting zeigen, dass sich aus dem Alter und dem Proteingehalt zusätzliche Information gewinnen lässt (s. Tabelle 18 und 19). Relativ betrachtet sind IPF-Patienten im Allgemeinen etwas älter. Die Klassifikationstabellen 20-22 zeigen schematisch wie die

Eingabewerte	N	err	CV	5-CV	$\widehat{\text{Err}}^{.632}$
1. Schritt					
X_A	448	0.538	0.547	0.551	0.543
X_{mw}	448	0.482	0.482	0.482	0.482
X_{PY}	436	0.573	0.573	0.574	0.589
X_{lP}	448	0.473	0.475	0.489	0.474
X_{lAlb}	446	0.478	0.478	0.48	0.485
X_{lZ}	445	0.51	0.51	0.51	0.516
X_{lM}	448	0.433	0.435	0.455	0.448
X_{lL}	448	**0.333**	**0.335**	**0.339**	**0.346**
X_{lG}	448	0.536	0.538	0.543	0.541
$X_{IVC\%}$	420	0.543	0.543	0.538	0.558
$X_{FEV1\%}$	427	0.585	0.607	0.606	0.591
X_{Tiff}	409	0.565	0.572	0.567	0.568
Stopp					
X_{lL}, X_A	448	**0.337**	0.344	0.35	0.347
X_{lL}, X_{mw}	448	0.344	0.35	0.357	0.348
X_{lL}, X_{PY}	436	0.346	0.349	0.346	0.35
X_{lL}, X_{lP}	448	0.346	0.353	0.346	0.355
X_{lL}, X_{lAlb}	446	0.343	0.35	0.359	0.358
X_{lL}, X_{lZ}	445	0.342	0.344	**0.344**	0.35
X_{lL}, X_{lM}	448	0.35	0.353	0.357	0.354
X_{lL}, X_{lG}	448	**0.337**	**0.342**	**0.344**	**0.345**
$X_{lL}, X_{IVC\%}$	**420**	0.336	*0.338*	0.338	*0.342*
$X_{lL}, X_{FEV1\%}$	**427**	0.347	0.349	0.351	0.35
X_{lL}, X_{Tiff}	**409**	*0.333*	0.34	*0.335*	0.344
X_{lL}, X_{lG}, X_{mw}	448	0.342	0.35	0.344	**0.342**
X_{lL}, X_{lG}, X_{lAlb}	446	**0.336**	**0.341**	**0.336**	0.347

Tabelle 15: LDA

Eingabewerte	N	err	CV	5-CV	$\widehat{\text{Err}}^{.632}$
1. Schritt					
X_A	448	0.54	0.54	0.554	0.544
X_{mw}	448	0.482	0.482	0.482	0.482
X_{PY}	436	0.583	0.583	0.583	0.589
X_{lP}	448	0.46	0.467	0.471	0.468
X_{lAlb}	446	0.487	0.491	0.489	0.487
X_{lZ}	445	0.51	0.51	0.512	0.52
X_{lM}	448	0.453	0.458	0.462	0.457
X_{lL}	448	**0.364**	**0.364**	**0.366**	**0.365**
X_{lG}	448	0.533	0.536	0.536	0.539
2. Schritt					
X_{lL}, X_{lG}	448	**0.346**	**0.35**	**0.353**	**0.35**
3. Schritt					
X_{lL}, X_{lG}, X_{mw}	448	0.33	**0.339**	**0.337**	**0.339**
4. Schritt					
$X_{lL}, X_{lG}, X_{mw}, X_A$	448	**0.266**	**0.295**	**0.326**	**0.298**
Stopp					

Tabelle 16: QDA

Eingabewerte	N	err	CV	5-CV	$\widehat{\text{Err}}^{.632}$
1. Schritt					
X_A	448	0.525	0.542	0.554	0.537
X_{mw}	448	0.482	0.482	0.482	0.482
X_{PY}	436	0.573	0.573	0.576	0.586
X_{lP}	448	0.458	0.467	0.462	0.471
X_{lAlb}	446	0.487	0.496	0.493	0.489
X_{lZ}	445	0.51	0.51	0.51	0.516
X_{lM}	448	0.442	0.446	0.44	0.445
X_{lL}	448	**0.333**	**0.333**	**0.339**	**0.342**
X_{lG}	448	0.538	0.538	0.54	0.542
Stopp					
X_{lL}, X_A	448	0.339	0.344	0.35	0.345
X_{lL}, X_{mw}	448	0.35	0.357	0.362	0.352
X_{lL}, X_{PY}	436	0.342	0.349	0.358	0.348
X_{lL}, X_{lP}	448	0.335	0.337	0.353	0.345
X_{lL}, X_{lAlb}	446	0.336	0.343	0.357	0.35
X_{lL}, X_{lZ}	445	0.335	0.342	0.342	0.344
X_{lL}, X_{lM}	448	**0.33**	**0.333**	**0.328**	**0.34**
X_{lL}, X_{lG}	448	0.339	0.346	0.346	0.342

Tabelle 17: Multinomiale Regression

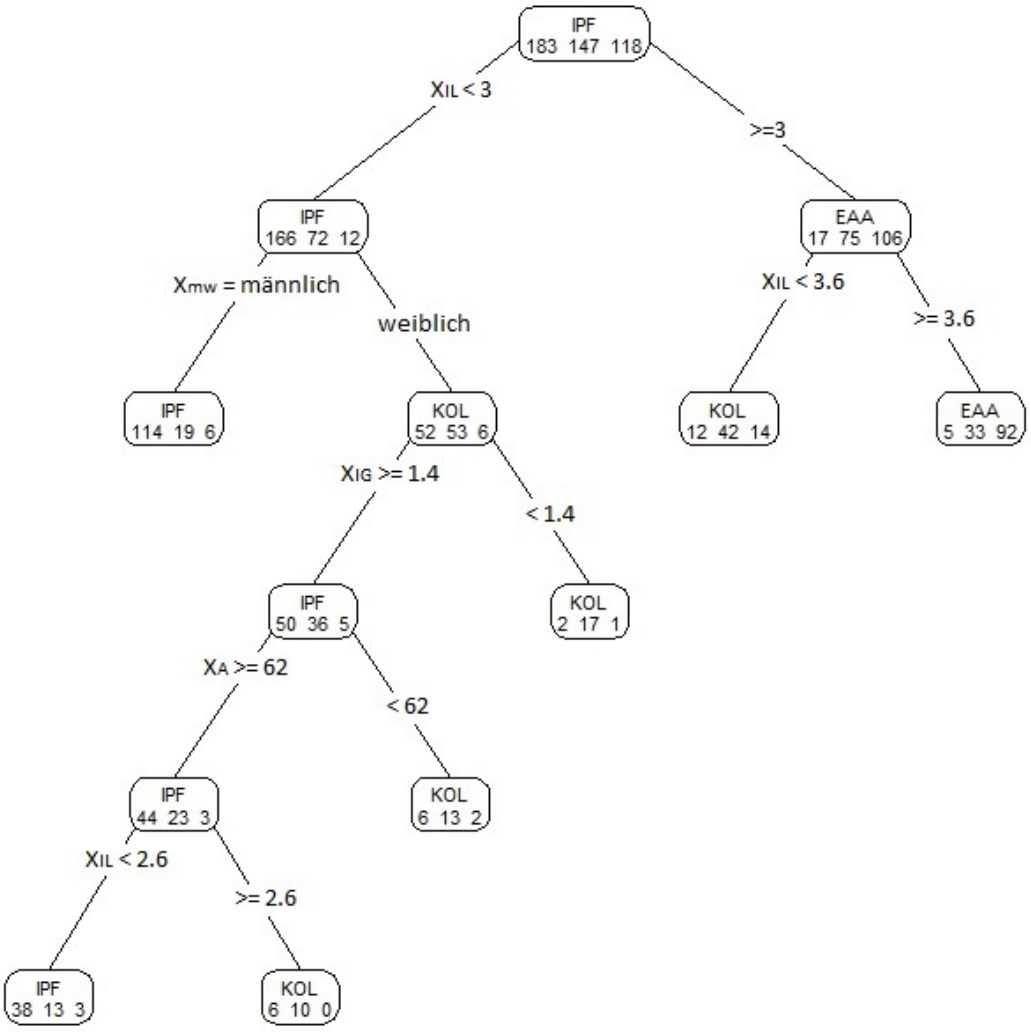

Abbildung 5: Klassifikationsbaum (Gini-Index)

X_{lL}	X_A	X_{mw}	X_{lM}	X_{lP}	X_{lG}	X_{lAlb}	X_{PY}	X_{lZ}
55.3	10.7	8.9	7.4	5.7	5.2	3.7	2	1.1

Tabelle 18: Relevanz bzgl. des Informationsgewinns in % (Bagging, B=30)

X_{lL}	X_A	X_{lP}	X_{lM}	X_{lG}	X_{lAlb}	X_{lZ}	X_{mw}	X_{PY}
25.02	15.99	15.13	9.88	9.86	8.79	6.45	4.79	4.11

Tabelle 19: Relevanz bzgl. des Informationsgewinns in % (Boosting mit SAMME, M=10)

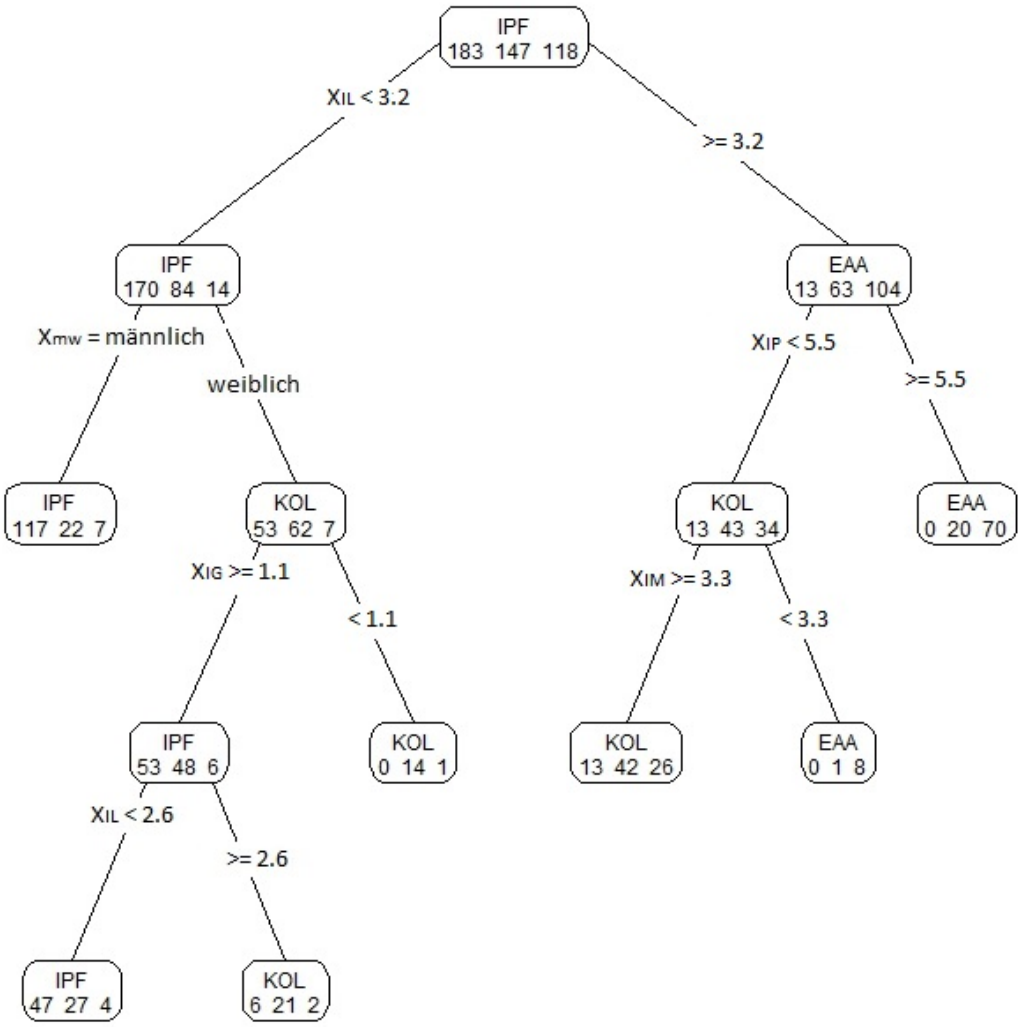

Abbildung 6: Klassifikationsbaum (Kreuzentropie)

	IPF	KOL	EAA
IPF	**159**	38	8
KOL	20	**93**	20
EAA	4	16	**90**

Tabelle 20: Bagging, Trainingsfehler 0.237, $B = 30$

	IPF	KOL	EAA
IPF	**175**	13	3
KOL	8	**124**	9
EAA	0	10	**106**

Tabelle 21: Boosting mit SAMME, Traingsfehler 0.096, $M = 10$

	IPF	KOL	EAA
IPF	**138**	29	7
KOL	41	**69**	31
EAA	9	29	**78**

Tabelle 22: Random Forests, Trainingsfehler 0.339, $B = 30$

Ensemble der verschiedenen Methoden abgestimmt haben. Sie verdeutlichen, dass die Hauptaufgabe der Kategorisierung darin besteht, die Gruppe IPF von KOL und KOL von EAA zu unterscheiden. Beim Boosting wurde in jedem Iterationsschritt ein Klassifikationsbaum an einen Bootstrap-Datensatz angepasst. Zudem wurde der Gini-Index zur Konstruktion von Bäumen bei allen Ensemble-Methoden angewendet. Verwendet man bei der Bagging- und Random-Forests-Methode $B = 60$ Bäume, so fällt der Trainingsfehler auf 0.192, bzw. auf 0.332. Der SAMME-Algorithmus lernt mit jedem Iterationsschritt mehr über den Datensatz bis er ihn mit einem Trainingsfehler von null Prozent vollständig korrekt klassifizieren kann (s. Tabelle 23). Der durch die Kreuzvalidierung geschätzte erwartete Testfehler sinkt jedoch nur bedingt, sodass es sich dabei um überangepasste Modelle handelt.

M	1	5	10	15	20	25	30	35	40
5-CV	0.39	0.40	0.40	0.37	0.38	0.35	0.37	0.33	0.36
err	0.30	0.19	0.08	0.02	0.01	0.00	0.00	0.00	0.00

Tabelle 23: Boosting mit SAMME abhängig von M

8.3 Information aus Lymphozyten und Proteinen

Kapitel 8.1 und 8.2 verdeutlichen, dass sich der Protein- und Albumingehalt durchaus für eine Kategorisierung der ILDs eignen. Sie sind jedoch mit dem Lymphozytenanteil korreliert, der klarer zwischen ihnen unterscheiden kann. Die Klassifikationsverfahren bevorzugen so die Lymphozyten und verzichten für eine geringere Modellkomplexität auf die Eingabewerte Protein und Albumin, die einen vergleichbaren Informationsgehalt besitzen. Dabei empfiehlt sich von diesen beiden wiederum der Proteingehalt, da seine Relevanz durch Boosting und Bagging höher bewertet wird und die entsprechenden Modelle einen niedrigeren Trainingsfehler aufweisen als die mit Albumin. Erfasst man nur den Albumingehalt in der BAL und nicht die übrigen Eiweiße, geht daher Information verloren.
Tabelle 24 listet die Trainingsfehler der Support-Vector-Machines, für die eine Hyperebene zur Trennung der jeweiligen Gruppen berechnet wurde, auf. Der Tuning-Parameter wurde dabei als $C = 1$ festgesetzt. Links unten wurden jeweils die Proteine als einziger Eingabewert verwendet, rechts oben Albumin. Die Trainingsfehler der Modelle mit Protein sind häufig etwas niedriger und verdeutlichen noch einmal den Vorteil des Gesamtproteingehalts.
Ein eindeutigeres Ergebnis liefert der Vergleich von Lymphozyten und Protein (s. Tabelle 25). Die Immunzellen sind der klare Sieger, während jedoch zur Trennung einiger Gruppen beide Eingabewerte zu denselben Trainingsfehlern führen. Die Überführung der Daten in einen höherdimensionalen Raum durch Verwendung radialer Basisfunktionen bedeuten kaum verbesserte Ergebnisse (s. Tabelle 26), sodass für einen geringeren Rechenaufwand und für eine bessere Interpretierbarkeit die lineare Trennung zu bevorzugen ist.

Gruppe	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
IPF	-	**0.234**	**0.381**	0.348	**0.2**	**0.099**	0.137	0.326
COP	0.254	-	**0.322**	0.237	0.388	0.177	0.135	0.204
KOL	0.4	0.325	-	0.391	0.322	**0.215**	0.166	0.333
RBILD	**0.346**	**0.201**	**0.344**	-	0.228	**0.131**	0.23	0.476
EAA	0.226	**0.382**	**0.291**	**0.181**	-	**0.253**	0.197	0.238
LIP	0.125	0.177	0.218	0.159	0.258	-	**0.304**	0.133
NSIP	0.137	**0.132**	**0.165**	0.23	0.197	0.314	-	0.214
KON	**0.295**	**0.176**	**0.298**	**0.422**	**0.155**	**0.101**	**0.188**	-

Tabelle 24: SVMs (linear) mit X_{lP}/X_{lAlb} links unten/rechts oben, $C = 1$

Gruppe	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
IPF	-	**0.155**	**0.264**	**0.346**	**0.096**	**0.054**	**0.108**	0.32
COP	0.254	-	0.343	**0.097**	0.382	0.177	0.132	**0.13**
KOL	0.4	**0.325**	-	**0.217**	**0.245**	0.218	0.165	**0.292**
RBILD	0.346	0.201	0.344	-	**0.056**	**0.022**	**0.079**	**0.383**
EAA	0.226	0.382	0.291	0.181	-	0.258	0.197	**0.064**
LIP	0.125	0.177	0.218	0.159	0.258	-	**0.257**	**0.024**
NSIP	0.137	0.132	0.165	0.23	0.197	0.314	-	**0.087**
KON	**0.295**	0.176	0.298	0.422	0.155	0.101	0.188	-

Tabelle 25: SVMs (linear) mit X_{lP}/X_{lL} links unten/rechts oben, $C = 1$

Gruppe	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
IPF	-	**0.155**	**0.27**	**0.336**	**0.103**	**0.058**	**0.108**	0.312
COP	0.254	-	0.346	**0.101**	0.382	0.177	0.132	**0.134**
KOL	0.364	**0.334**	-	**0.201**	**0.226**	0.218	0.165	**0.288**
RBILD	0.321	0.205	0.348	-	**0.051**	**0.014**	**0.087**	**0.366**
EAA	0.223	0.382	0.306	0.181	-	0.258	0.197	0.064
LIP	0.121	0.177	**0.197**	0.145	0.258	-	0.243	0.024
NSIP	0.137	0.132	0.165	0.198	0.197	0.3	-	0.087
KON	**0.292**	0.161	0.315	0.427	0.155	0.101	0.188	-

Tabelle 26: SVMs (radial) mit X_{lP}/X_{lL} links unten/rechts oben, $C = 1$, $\sigma^2 = 1$

9 Test auf IPF

Im diagnostischen Alltag ist es besonders wichtig, Indizien für und gegen die IPF zu finden, da sie eine der häufigsten interstitiellen Lungenerkrankungen mit hoher Mortalität ist. In der medizinischen Literatur gilt es als akzeptiert, dass eine Lymphozytose, d. h. ein erhöhter Lymphozytenanteil von mehr als 40%, eine IPF weitestgehend ausschließt. Diese Tatsache bestätigt sich auch in dem vorliegenden Datensatz. Nur 2 der 183 BAL-Befunde der IPF-Patienten zeigt eine Lymphozytose. Geht man von Beta-binomialverteilten Lymphozyten mit den Parametern $\hat{\alpha} = 1.282694$ und $\hat{\beta} = 12.57906$ aus, so kann die Hypothese, dass ein Patient an IPF leidet mit den kritischen Werten aus Tabelle 27 verworfen werden. Auch ein stark erhöhter Proteingehalt kann diese Hypothese widerlegen. Unter der Voraussetzung $X_{lL} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ mit $\hat{\mu} = 4.864574$, $\hat{\sigma}^2 = 0.1793625$ sind die entsprechenden Werte ebenfalls in Tabelle 27 angegeben.

Quantil	X_L	X_{lP}	X_P
0.90	19.7%	5.407	223.0
0.95	24.3%	5.561	260.1
0.99	33.9%	5.850	347.2
0.999	45.3%	6.173	479.8

Tabelle 27: Quantile unter $10X_{lL} \sim \text{BeB}(1000, \hat{\alpha}, \hat{\beta})$ und $X_{lP} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$

90%-Niveau	N	$X_L \geq 19.7\%$	$X_{lP} \geq 5.407$	$X_L \geq 19.7\%$ und $X_{lP} \geq 5.407$
IPF	183	17	17	1
andere	711	390	254	222
95%-Niveau	N	$X_L \geq 24.3\%$	$X_{lP} \geq 5.561$	$X_L \geq 24.3\%$ und $X_{lP} \geq 5.561$
IPF	183	13	10	0
andere	711	352	219	189
99%-Niveau	N	$X_L \geq 33.9\%$	$X_{lP} \geq 5.850$	$X_L \geq 33.9\%$ und $X_{lP} \geq 5.850$
IPF	183	6	3	0
andere	711	289	186	144
99.9%-Niveau	N	$X_L \geq 45.3\%$	$X_{lP} \geq 6.173$	$X_L \geq 45.3\%$ und $X_{lP} \geq 6.173$
IPF	183	2	1	0
andere	711	210	132	90

Tabelle 28: Empirische Auswertung in absoluten Häufigkeiten

Gruppe	IPF	COP	KOL	RBILD	EAA	LIP	NSIP	KON
$\hat{\rho}(X_P, X_L)$	0.07	0.28	0.35	0.76	0.52	0.51	0.25	0.26
$\hat{\rho}_s(X_P, X_L)$	0.19	0.37	0.50	0.06	0.65	0.51	0.32	0.23

Tabelle 29: Pearson- bzw. Spearman-Korrelationskoeffizienten ($\hat{\rho}$ bzw. $\hat{\rho}_s$)

Eine interessante Entdeckung ist, dass die beiden Entzündungswerte in der Gruppe IPF schwächer miteinander korrelieren (s. Tabelle 29) als in anderen. IPF-Patienten mit Ausreißern in den Lymphozytendaten haben nach Tabelle 28 im Allgemeinen keinen extrem hohen Proteingehalt. Das Protein in der BAL ist also ein wichtiges zusätzliches Hilfsmittel um eine IPF entschiedener auszuschließen, falls beide Werte erhöht sind.

10 Fazit

In der medizinischen Literatur (s. American Thoracic Society und European Respiratory Society [2]) gilt es als akzeptiert, dass die Anzahl der Immunzellen in der BAL asymmetrisch verteilt ist. Darüber hinaus wurde in Kapitel 7.3 gezeigt, dass die Beta-Binomialverteilung eine plausible Parametrisierung dieser Verteilung ist. Für zukünftige statistische Auswertungen des BAL-Differentialbilds ist es wichtig, die Ergebnisse in absoluten Zahlen festzuhalten, damit eine weitere Transformation der Daten obsolet wird. Mit Hilfe dieser Parametrisierung ist es möglich den Zusammenhang zwischen der BAL und den ILDs zu untersuchen.
Die Anwendung der Klassifikationsverfahren auf den Schildge-Datensatz zeigt, dass sich das Protein als Eingabewert besser eignet als das Albumin (s. Kapitel 8). Sofern es daher kein plausibles medizinisches Gegenargument gibt, ist der Gesamtproteingehalt der Teilmenge des Albumingehalts vorzuziehen. Für eine Kategorisierung ist der Lymphozytenanteil die unangefochten wichtigste Variable. Dennoch ist der Proteingehalt ein wichtiges Indiz, da die Informationslage beim Diagnostizieren einer ILD dünn ist. Erhöhte Eiweißwerte sprechen gegen eine IPF (s. Kapitel 9).
Für einen theoretischen Vergleich der Klassifikationsmethoden sind Definitionen für Varianz und Bias kategorialer Zufallsvariablen unerlässlich, um den entstehenden Testfehler zu analysieren. Diese Vorgehensweise hat noch nicht den Weg in die Standardliteratur des maschinellen Lernens gefunden. Die Ausführungen in Kapitel 3 weisen auf Vorteile hin, wenn dabei mit Notation und Sprechweisen zwischen Klassifikationsregeln und -methoden unterschieden wird. Durch die verschiedenen Effekte lässt sich dann die Zusammensetzung des (erwarteten) Testfehlers untersuchen. Hastie, Tibshirani und Friedman [7, S. 312] bemerken, dass Klassifikationsbäume eine unter ihrem Verständis hohe Varianz aufweisen, da kleine Veränderungen im Trainingsdatensatz die Entscheidungsregel entscheidend verändern können. Den Grund hierfür sehen sie in der hierarchischen Struktur der Klassifikationsbäume. Wie auch James [8] argumentieren sie, dass der Erfolg des Baggings auf einer Reduktion der Varianz beruht. Die jeweiligen Definitionen von Varianz sind dabei kontrovers. Um die in Kapitel 3.2 vorgeschlagenen Definitionen des methodenbedingten Effekts zu validieren, wäre es im Rahmen einer Simulationsstudie von Interesse, zu untersuchen, ob dieser bei CART-Bäumen höher ist als bei anderen Methoden und ob das Bagging ihn reduziert.

11 Anhang

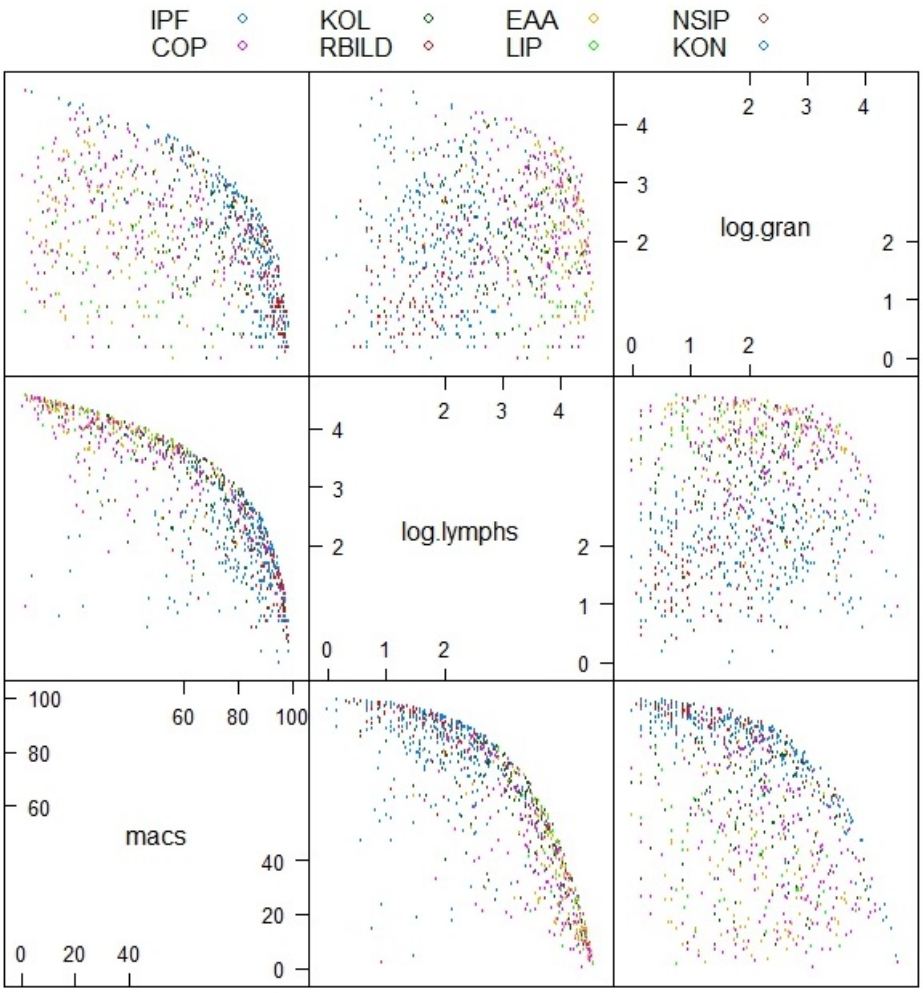

Abbildung 7: Systematischer Zusammenhang der BAL-Parameter

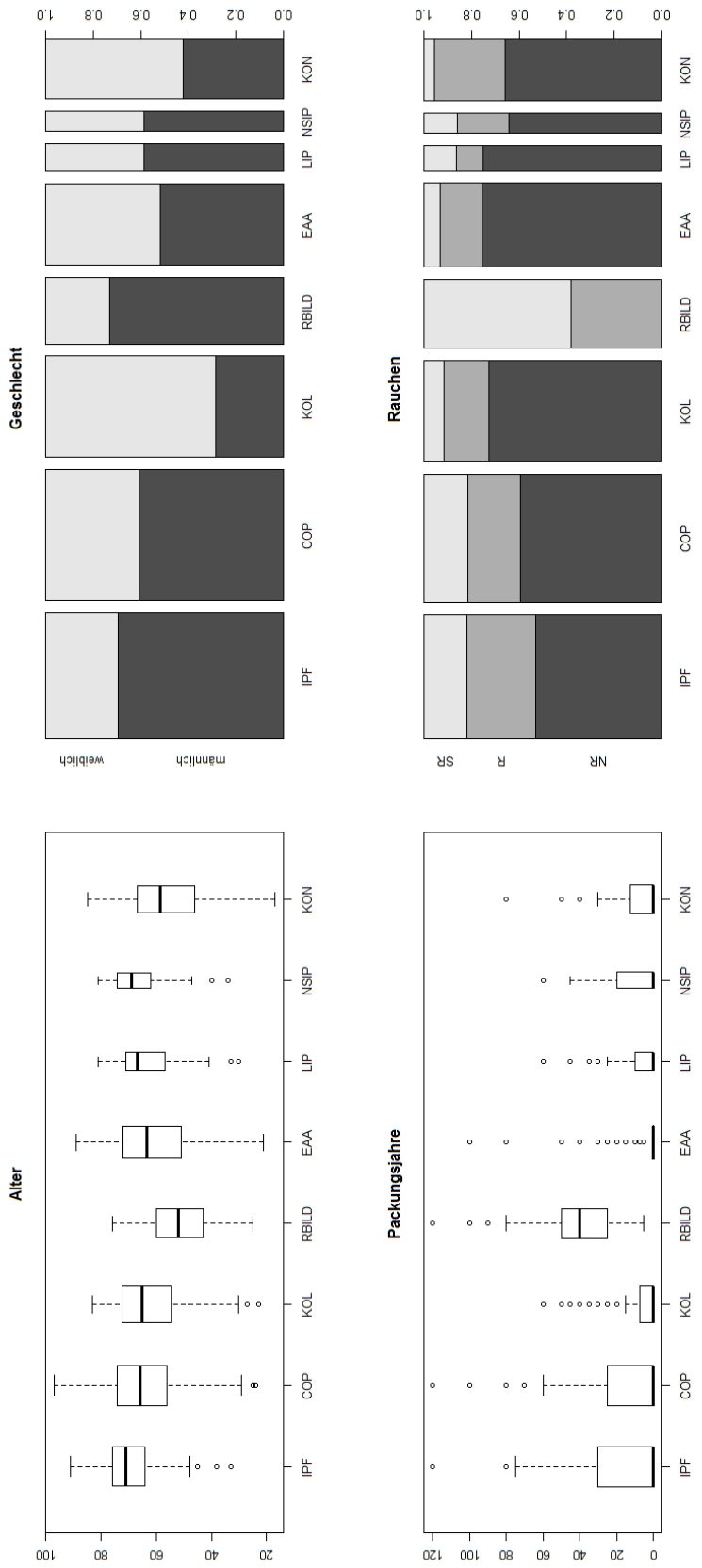

Abbildung 8: Epidemiologie (NR: Nichtraucher 0 PY, R: Raucher 0-30 PY, SR: Starker Raucher mehr als 30 PY)

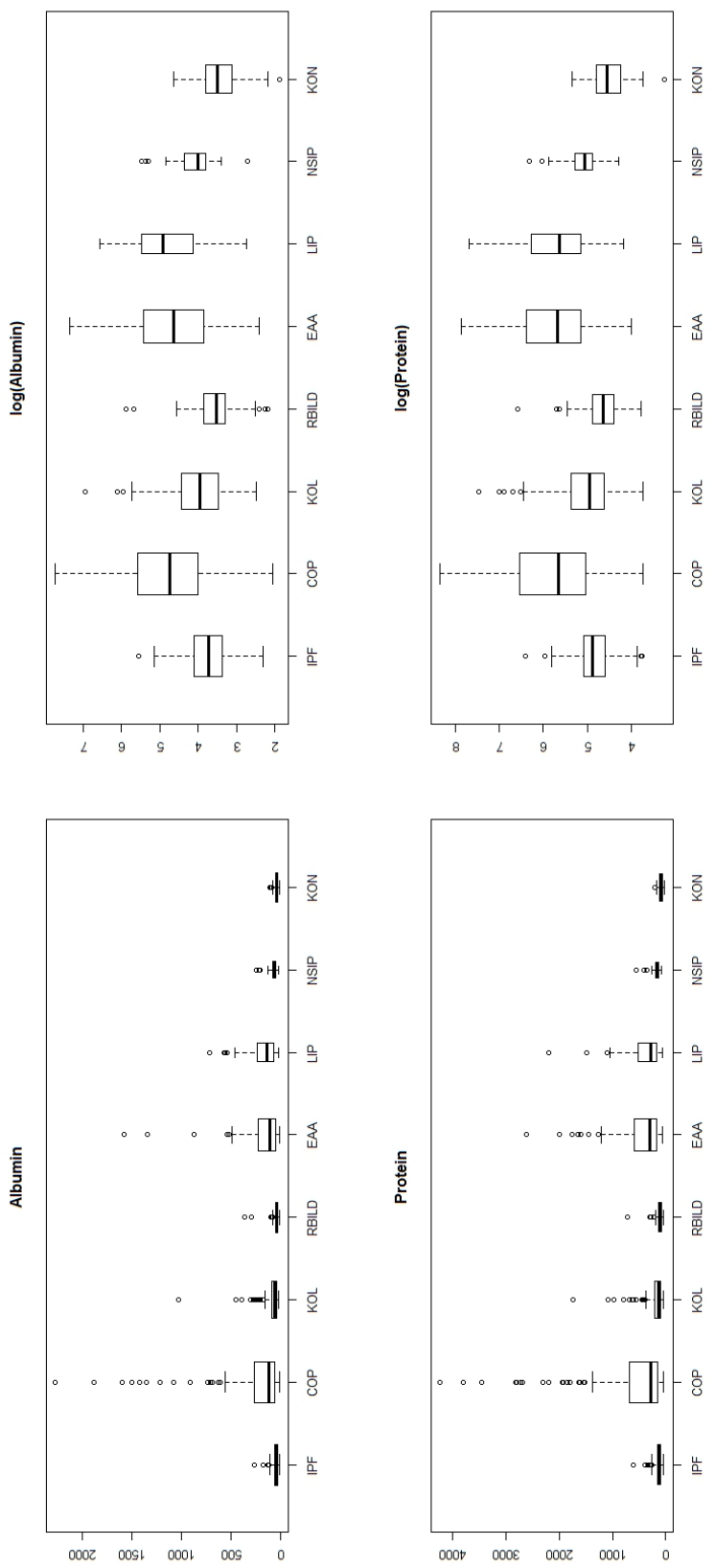

Abbildung 9: BAL I

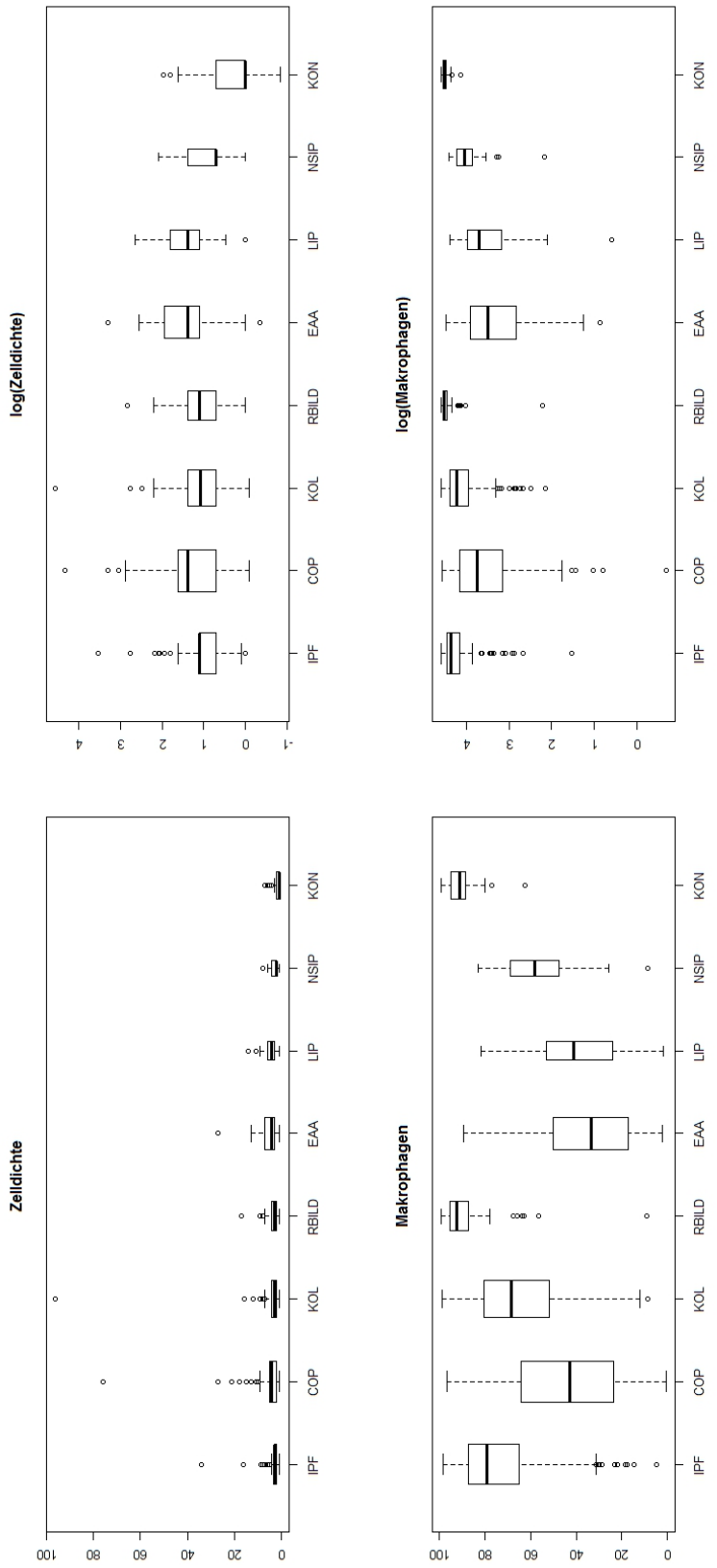

Abbildung 10: BAL II

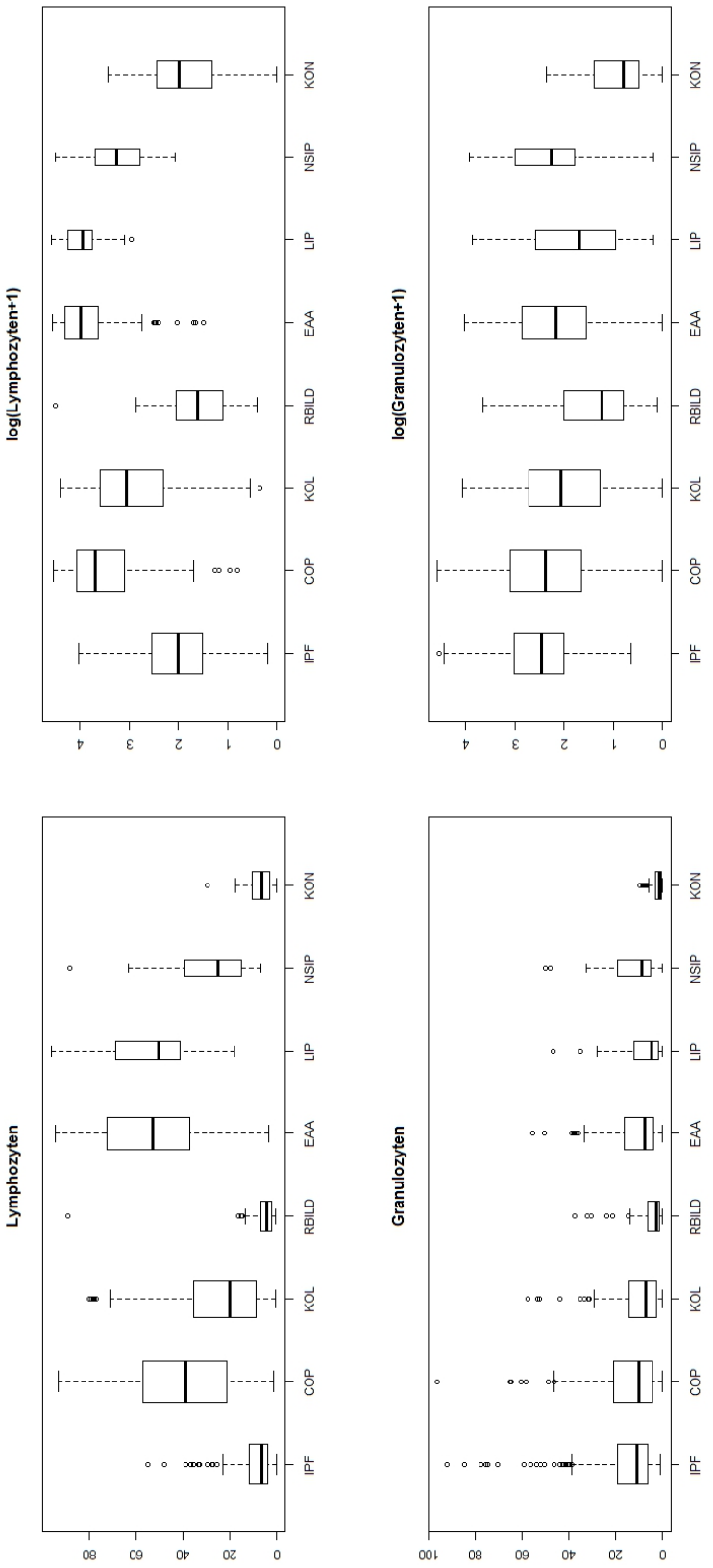

Abbildung 11: BAL III

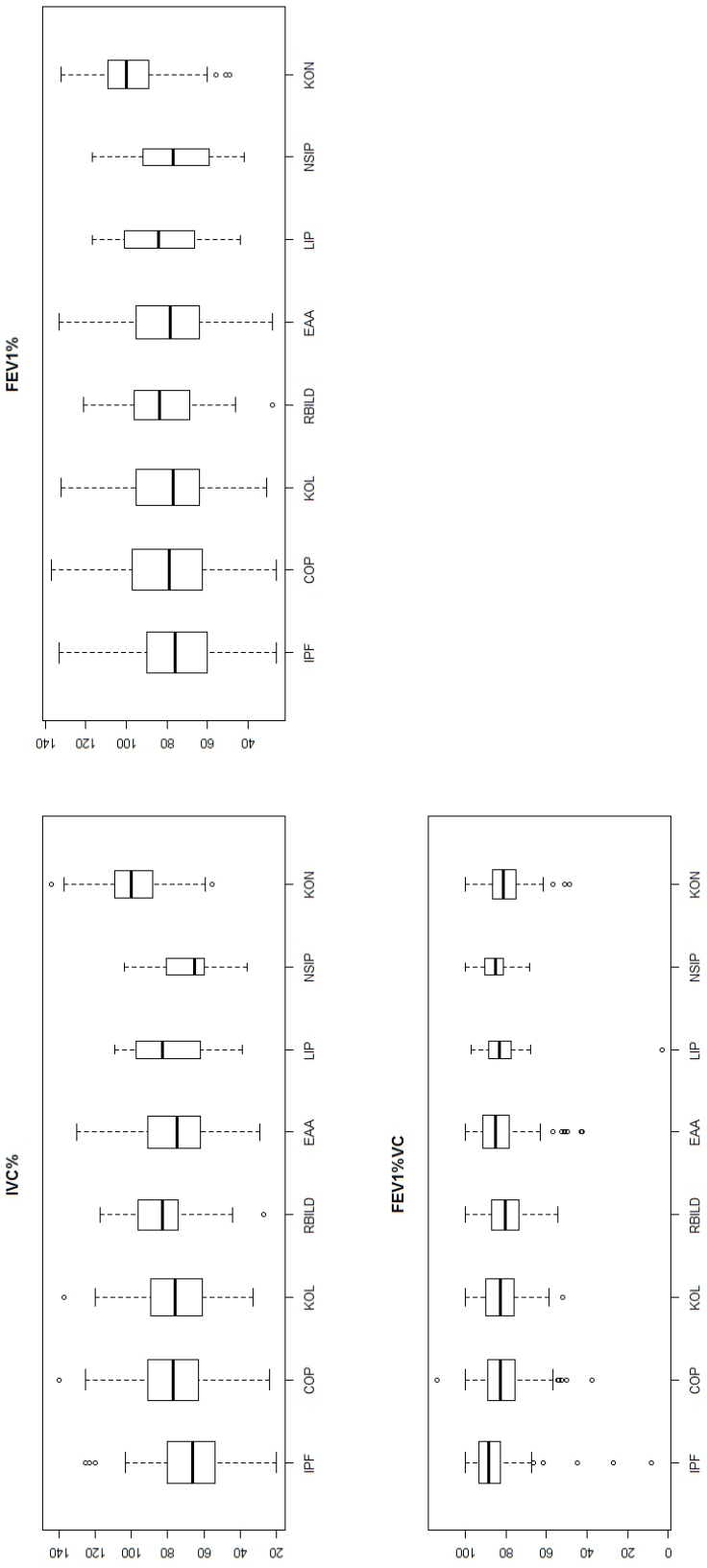

Abbildung 12: Spirometrie

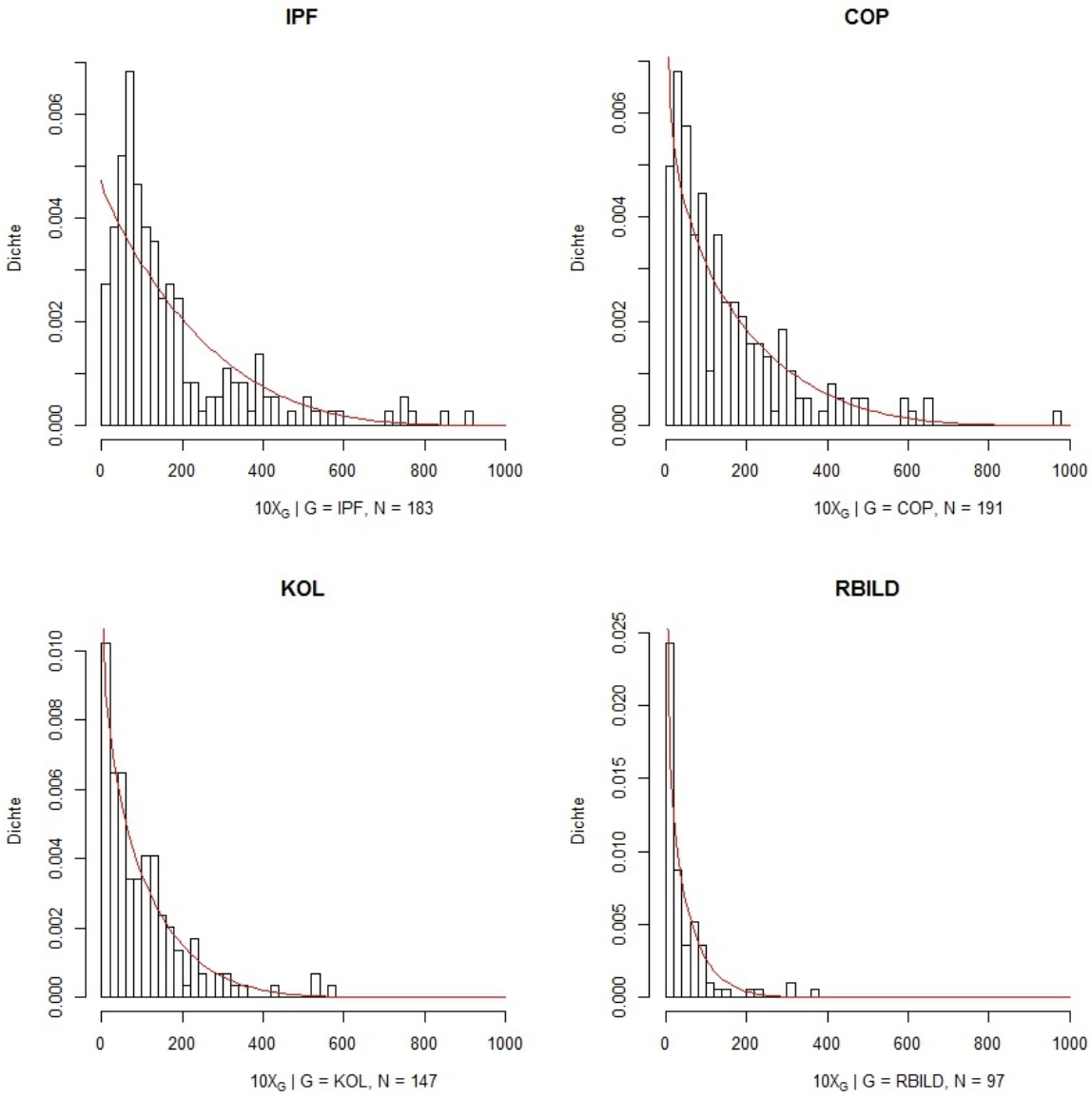

Abbildung 13: Beta-Binomialverteilung der Granulozytendaten I

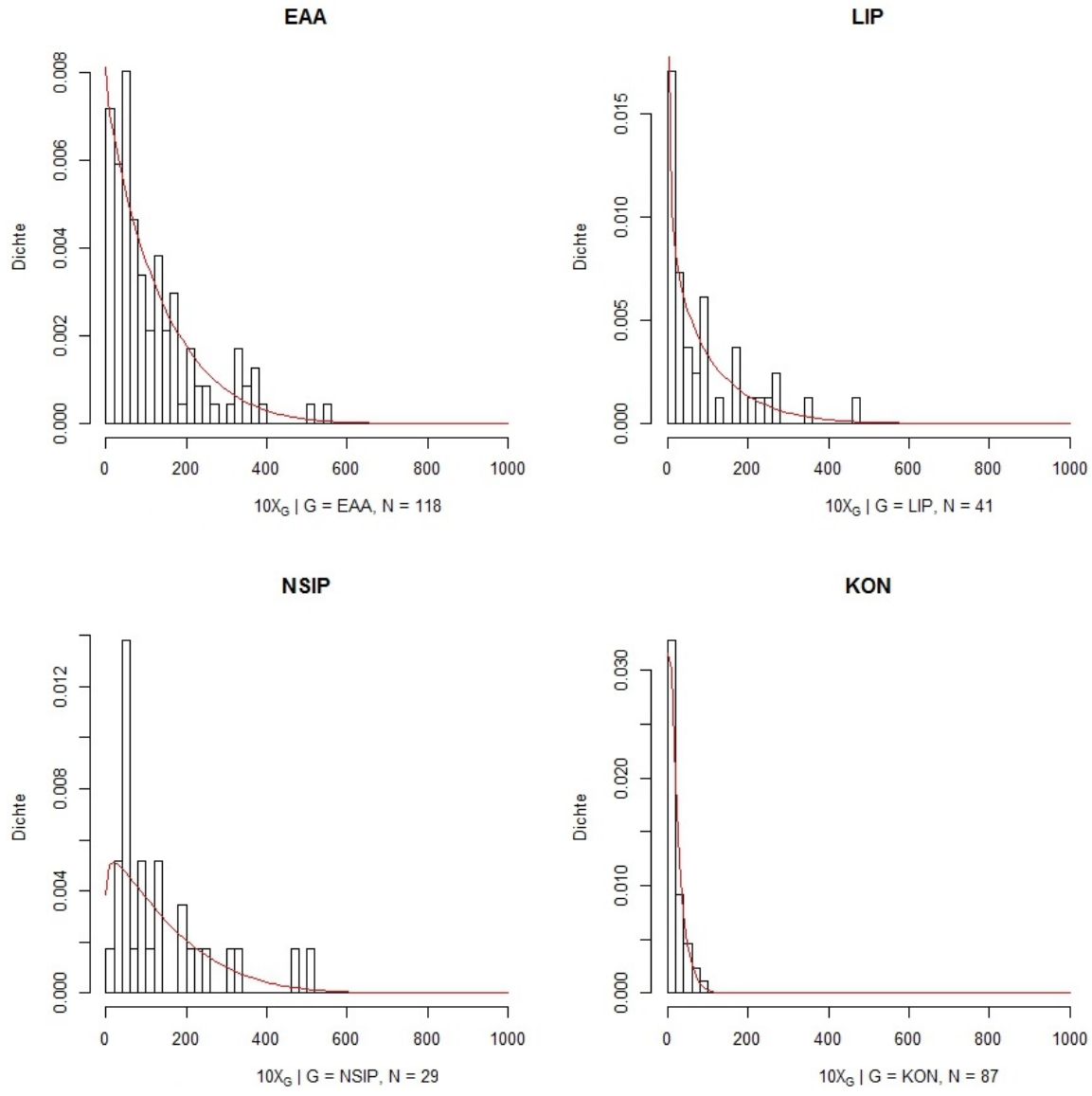

Abbildung 14: Beta-Binomialverteilung der Granulozytendaten II

$\hat{\rho}(X_i, X_j)$	X_A	X_{PY}	X_P	X_{Alb}	X_Z	X_M	X_L	X_G	$X_{IVC\%}$	$X_{FEV1\%}$	X_{Tiff}
X_A	1.00	-0.06	-0.08	-0.08	-0.09	-0.07	0.02	0.11	-0.10	-0.01	-0.02
X_{PY}	-0.06	1.00	-0.08	-0.07	-0.03	0.19	-0.18	-0.05	-0.02	-0.07	-0.19
X_P	-0.08	-0.08	1.00	**0.91**	**0.34**	**-0.49**	**0.49**	0.11	-0.04	-0.05	0.01
X_{Alb}	-0.08	-0.07	**0.91**	1.00	**0.33**	**-0.43**	**0.45**	0.06	-0.03	-0.03	0.02
X_Z	-0.09	-0.03	**0.34**	**0.33**	1.00	**-0.34**	0.25	0.23	-0.04	-0.02	0.04
X_M	-0.07	0.19	**-0.49**	**-0.43**	**-0.34**	1.00	**-0.87**	**-0.44**	0.13	0.12	-0.03
X_L	0.02	-0.18	**0.49**	**0.45**	0.25	**-0.87**	1.00	-0.06	-0.04	-0.03	0.03
X_G	0.11	-0.05	0.11	0.06	0.23	**-0.44**	-0.06	1.00	-0.20	-0.19	-0.01
$X_{IVC\%}$	-0.10	-0.02	-0.04	-0.03	-0.04	0.13	-0.04	-0.20	1.00	**0.87**	-0.23
$X_{FEV1\%}$	-0.01	-0.07	-0.05	-0.03	-0.02	0.12	-0.03	-0.19	**0.87**	1.00	0.16
X_{Tiff}	-0.02	-0.19	0.01	0.02	0.04	-0.03	0.03	-0.01	-0.23	0.16	1.00

Tabelle 30: Empirische Pearson-Korrelationskoeffizienten

Literatur

[1] Alfaro, E., Gamez, M. und Garcia, N., *adabag: An R Package for Classification with Boosting and Bagging*, Journal of Statistical Software, Volume 54, 2, S. 1–35, 2013.

[2] American Thoracic Society, European Respiratory Society, *American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias*, American Journal of Respiratory and Critical Care Medicine, Vol. 165, No. 2, 2002.

[3] Breiman, L., *Random Forests*, Machine Learning, Volume 45, Issue 1, S. 5-32, 2001.

[4] Breiman, L., Friedman, J., Olshen, R. und Stone, C., *Classification and Regression Trees*, Wadsworth, New York, 1984.

[5] Diestel, R., *Graphentheorie*, 2. Auflage, Springer-Verlag, Heidelberg, 2000.

[6] Domingos, P., *A Unified Bias-Variance Decomposition and its Applications* in *Proceedings of the 17th International Conference on Machine Learning*, S. 231-238, 2000.

[7] Hastie, T., Tibshirani, R. und Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. Auflage, Springer-Verlag, New York, 2009.

[8] James, G., *Variance and Bias for General Loss Functions*, Machine Learning, Volume 51, S. 115-135, 2003.

[9] James, G., Witten, D., Hastie, T. und Tibshirani, R., *An Introduction to Statistical Learning with applications in R*, 1. Auflage, Springer-Verlag, New York, 2013.

[10] Johnson, N. L., Kemp, A. W. und Kotz S., *Univariate Discrete Distributions*, 3. Auflage, John Wiley & Sons, 2005.

[11] Kroegel, C. und Bonella, F., *Klinische Pneumologie: das Referenzwerk für Klinik und Praxis*, 1. Auflage, Thieme, Stuttgart, 2014.

[12] Müller-Quernheim, J., *Interstitielle Lungenerkrankungen: Standards in Klinik, Diagnostik und Therapie*, 1. Auflage, Stuttgart, New York, Thieme, 2003.

[13] Murphy, K. M., Travers, P. und Walport, M., *Janeway Immunologie*, 7. Auflage, Spektrum Akademischer Verlag, Heidelberg, 2009.

[14] Meyer, K. C., Raghu, G., Baughman, R. P. et al., *An Official American Thoracic Society Clinical Practice Guideline: The Clinical Utility of Bronchoalveolar Lavage Cellular Analysis in Interstitial Lung Disease*, American Journal of Respiratory and Critical Care Medicine Volume 185, 2012.

[15] Schölkopf, B. und Smola, A. J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge (Massachusetts), London, 2002.

[16] Ulmer, W. T., Nolte, D., Lecheler, J. und Schäfer, T., *Die Lungenfunktion: Methodik und klinische Anwendungen*, 7. Auflage, Thieme, Stuttgart, 2003.

[17] Quanjer, Ph. H., Tammeling, G. J., Cotes, J. E., Pedersen, O. F., Peslin, R. und Yernault, J-C., *Lung volumes and forced ventilatory flows. Report Working Party Standardization of Lung Function Tests, European Community for Steel and Coal. Official Statement of the European Respiratory Society.*, Eur. Respir. J. 6, Suppl. 16, 1993.

[18] Zhu, J., Rosset, S., Zou, H. und Hastie, T., *Multi-class AdaBoost*, Technical Report 430, Department of Statistics, University of Michigan, 2006.

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen, als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe.

Karlsruhe, den 29. Oktober 2015