Modellvergleich, Modellauswahl und das Einstellen von Hyperparametern und Bestrafungstermen

Volker Tresp

Empirischer Modellvergleich

Testfehler (Generalisierungsfehler)

- Gegeben zwei Modellansätze M_1 und M_2, so wollen wir nachweisen, dass M_1 bessere Performanz besitzt als M_2
- Beispiel: M_1 bezeichnet ein Neuronales Netz und M_2 ist lineare Regression
- Wie schon mehrfach erwähnt, ist hierzu der Vergleich der Performanz auf den Trainingsdaten nicht verlässlich
- Von Interesse ist der erwartete Generalisierungsfehler, das heißt die Performanz auf neuen Daten
- Sei $L[y, f(\mathbf{x}, M_1)]$ eine Verlustfunktion (z.B. quadratischer Fehler, Klassifikationsfehler), dann ist von Interesse

$$E_{P(\mathbf{x},y)} L[y, f(\mathbf{x}, M_i)]$$

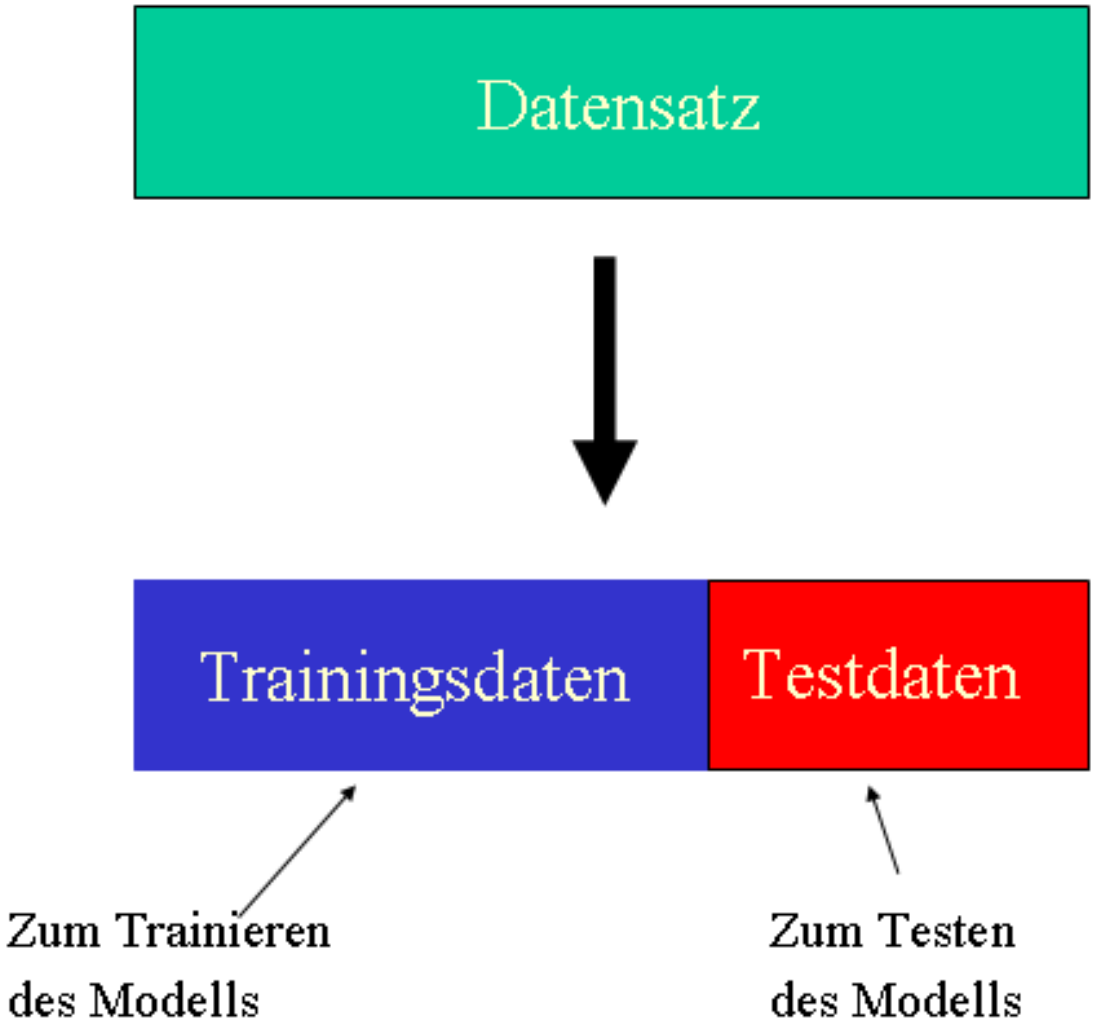
Empirische Approximation des Testfehlers

- Um diesen Ausdruck zu approximieren teilt man den vorhandenen Datensatz auf in einen Trainingsdatensatz und einen Generalisierungsdatensatz (Validierungssatz)
- Die Modelle werden nur auf den Trainingsdaten trainiert
- Ein erwartungstreuer Schätzer des Generalisierungsfehlers ist

$$E_{P(\mathbf{x},y)} L[y, f(\mathbf{x}, M_i)] \approx$$

$$J^{test}(M_i) = \frac{1}{|TEST|} \sum_{i \in TEST} L[y_i, f(\mathbf{x}_i, M_i)]$$

also einfach der mittlere Fehler auf den Testdaten


Datensatz
Trainingsdaten
Testdaten
Zum Trainieren des Modells
Zum Testen des Modells

Kreuzvalidierung

- Der Testfehler ist erwartungstreu, jedoch besitzt er oft erhebliche Varianz
- Daher ist der Modellansatz mit dem besseren Testfehler nicht notwendigerweise das bessere Verfahren
- Ein sichereres Verfahren ist die K-fache Kreuzvalidierung; typische Zahlen sind $K = 5$ oder $K = 10$
- Die Daten werden in K gleichgroße Gruppen partitioniert
- Für $j = 1, \ldots, K$: Die $j-$te Menge ist der Testdatensatz und die übrigen Datensätze agieren als Trainingsdaten

Kreuzvalidierung (2)

- So erhält man für jeden Modellansatz i nicht nur einen sondern K Testfehler $J^{test}(M_i, j)$
- Man kann nun den mittleren Testfehler berechnen als

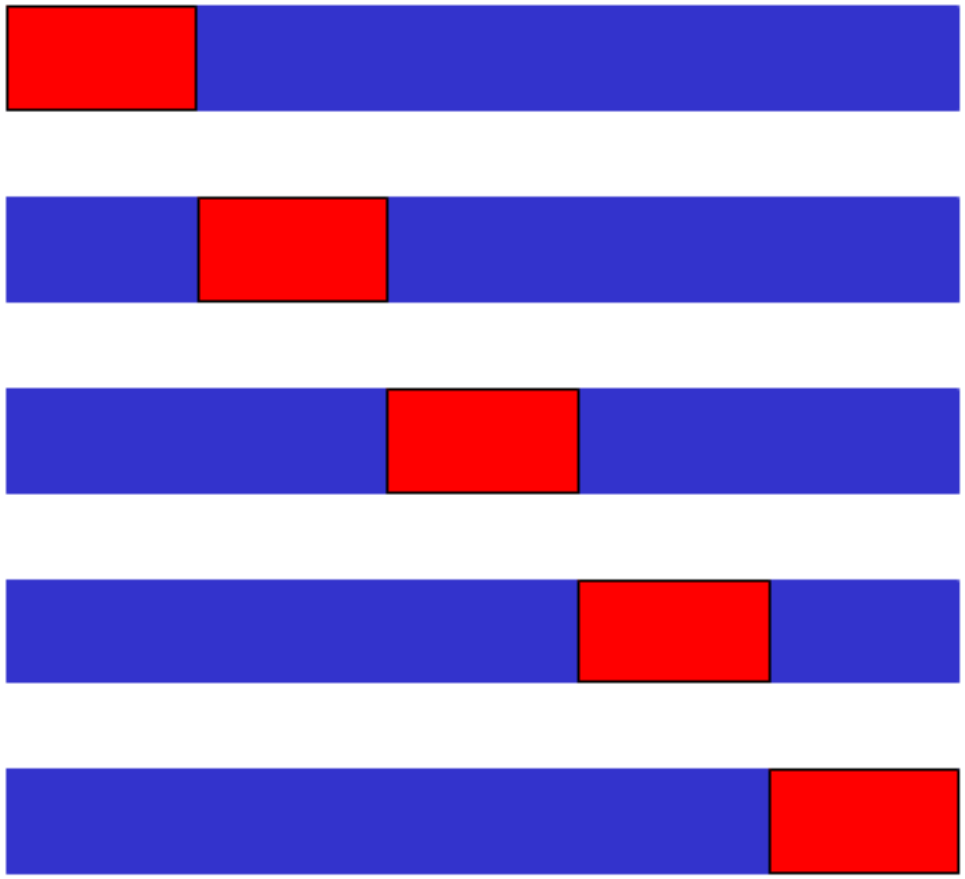
$$mean_i = \frac{1}{K} \sum_{j=1}^{K} J^{test}(M_i, j)$$

- Die Varianz des mittleren Testfehlers kann geschätzte werden zu

$$var_i = \frac{1}{K(K-1)} \sum_{j=1}^{K} (J^{test}(M_i, j) - mean_i)^2$$

- Man würde Modellansatz M_i als besser als M_j einstufen, wenn sich die Standardabweichungen nicht überlappen, das heisst, falls

$$mean_i + \sqrt{var_i} < mean_j - \sqrt{var_j}$$



5-Fache
Kreuzvalidierung:
Blau: Trainingsdaten
Rot: Testdaten

Gepaarte Tests

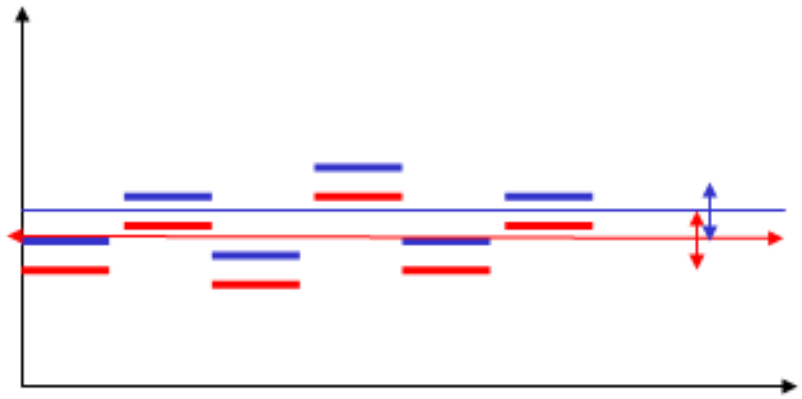
- Wenn man sehr wenige Daten hat, ist die Kreuzvalidierung manchmal nicht scharf genug
- Die Grundidee: nehmen wir an $K = 10$; wenn nun M_1 in neun der zehn Tests besser abschneidet als M_2, dann spricht dies stark für M_1
- Man berechnet die mittlere Differenz der Verfahren

$$\mathsf{MeanDiff}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} J^{test}(M_i, k) - J^{test}(M_j, k)$$

und analysiert, ob diese Differenz signifikant positiv (oder negativ ist); eine sorgfältigere Anlayse führt zum gepaarten T-Test (paired t-test)

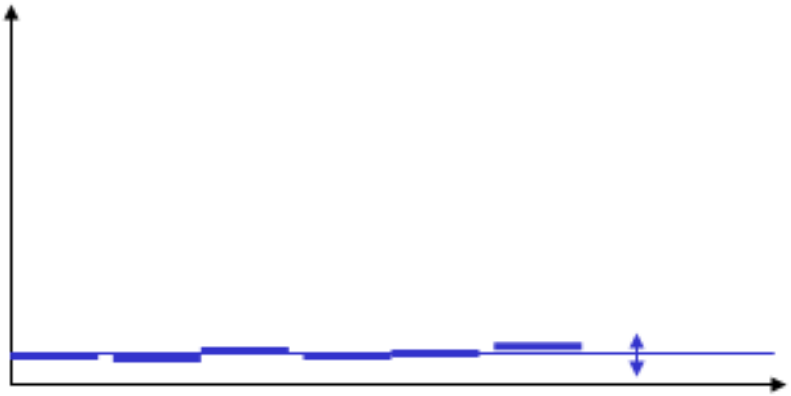
Testfehler für Schätzung eines Mittelwertes

Testfehler



- Basierend auf Mittelwert und Varianz kann nicht entschieden werde, dass Modellansatz 1(blau) signifikant besser ist als Modellansatz 2

Differenz der Testfehler



- Untersucht man jedoch die Differenz der Performanz ist die bessere Performanz von Modellansatz 1 (blau) signifikant

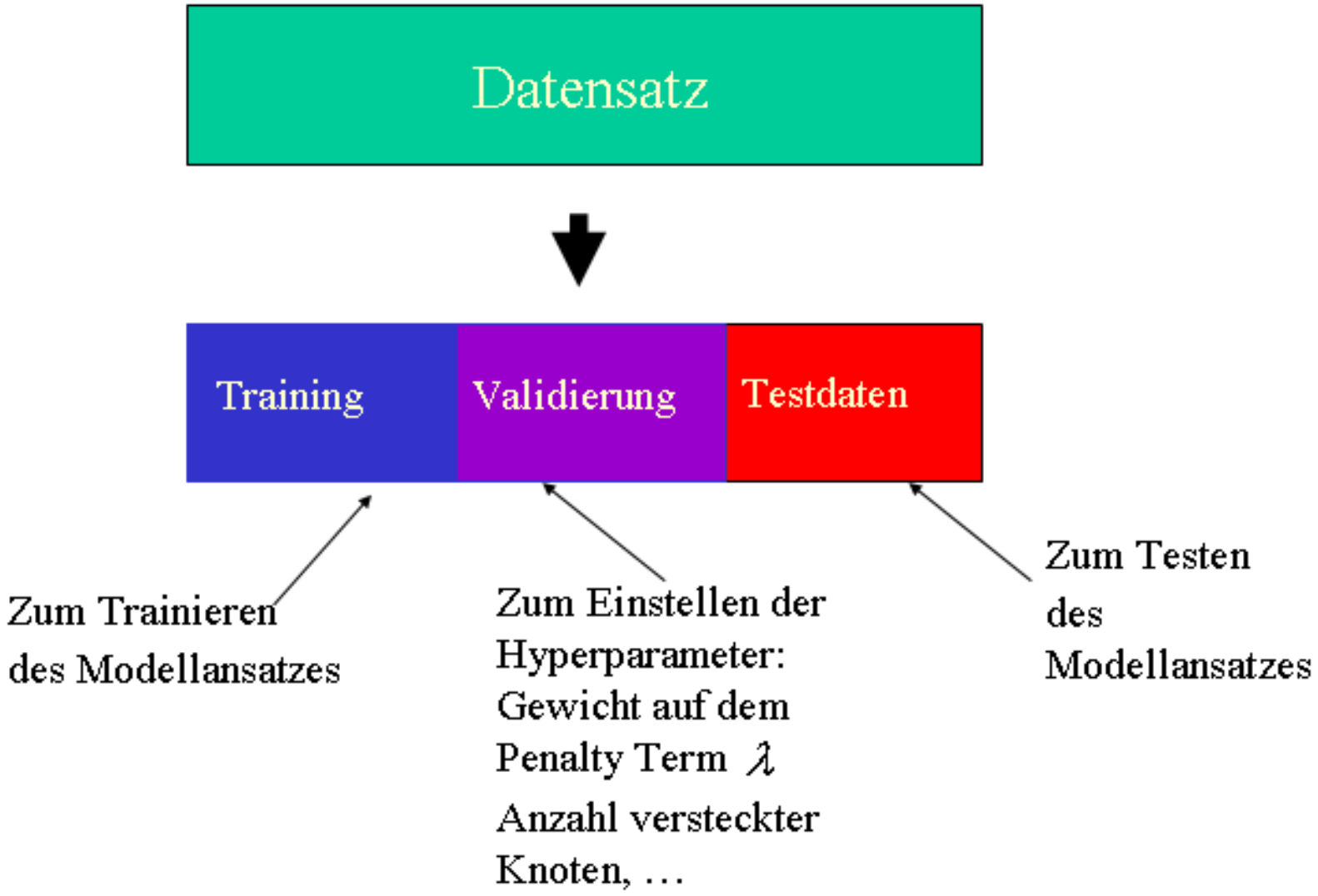
Empirische Einstellung der Hyperparameter

Hyperparameter

- Neben den eigentlichen Parametern, die durch den Lernprozess bestimmt werden, gibt es auch sogenannte Hyperparameter: typischerweise sind dies die Gewichtungen auf den Straftermen λ, die Anzahl der versteckten Knoten eines Neuronalen Netzes, ...
- Bayes'sche Verfahren haben hier einen Vorteil, da Hyperparameter einfach nur weitere Parameter im Modell darstellen, über die intergriert werden muss
- Die meisten anderen Verfahren tun sich schwerer mit einer prinzipiellen Bestimmung der Hyperparameter; eine universelle Lösung stellt die empirische Bestimmung dar

Hyperparameter(2)

- Die Idee ist eine drei-Einteilung der Daten in Trainings-, Validierungs-, und Testdaten
  - Das Modell wird auf den Trainingsdaten mit verschiedenen Werten der Hyperparameter trainiert
  - Es wird das Modell mit den entsprechenden Hyperparametern ausgewählt, welches auf den Validierungsdaten die beste Performanz gezeigt hat
  - Es wird der Testfehlers dieses optimierten Modells berechnet
- Ähnlich wie bei der Modellauswahl, kann natürlich auch die Bestimmung der Hyperparameter über Kreuzvalidierung erfolgen


Datensatz
Training
Validierung
Testdaten
Zum Trainieren des Modellansatzes
Zum Einstellen der Hyperparameter: Gewicht auf dem Penalty Term λ Anzahl versteckter Knoten, …
Zum Testen des Modellansatzes

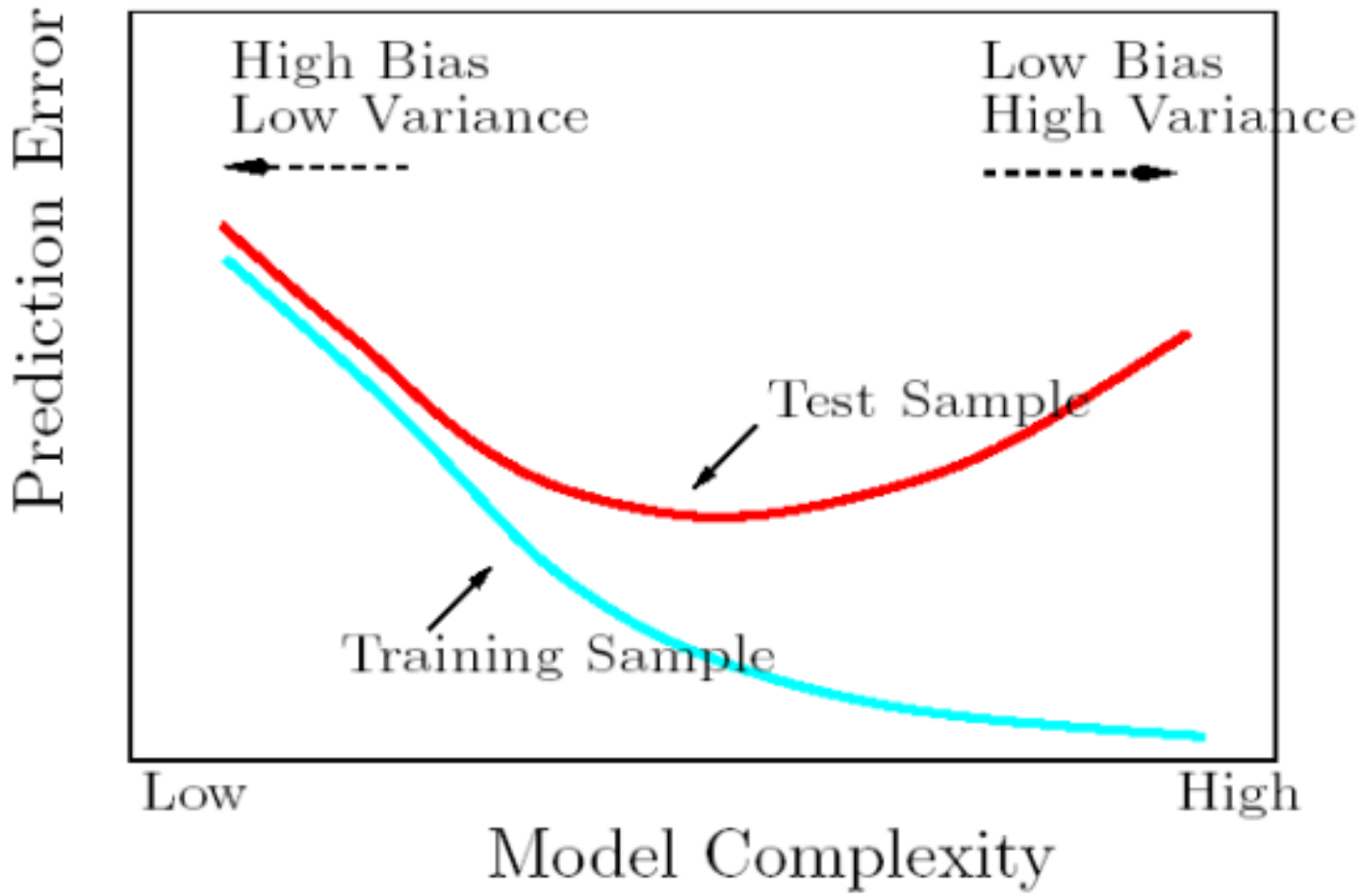
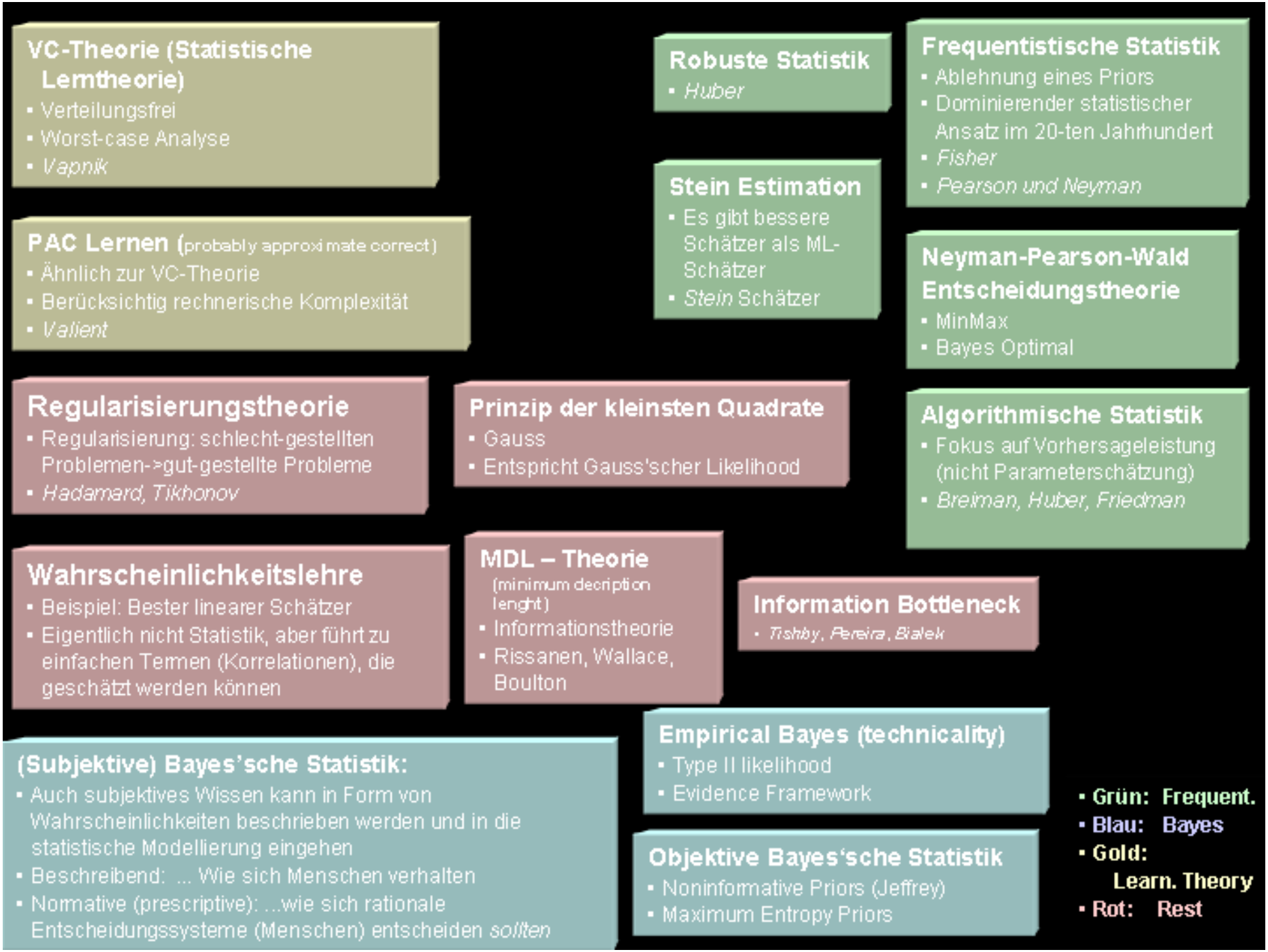


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

Lerntheorien und theoretische Abschätzungen des Generalisierungsfehlers

Überblick: Statistische Theorien und Lerntheorien


VC-Theorie (Statistische Lerntheorie)
• Verteilungsfrei
• Worst-case Analyse
• Vapnik
Robuste Statistik
• Huber
Frequentistische Statistik
• Ablehnung eines Priors
• Dominierender statistischer Ansatz im 20-ten Jahrhundert
• Fisher
• Pearson und Neyman
Stein Estimation
• Es gibt bessere Schätzer als ML-Schätzer
• Stein Schätzer
PAC Lernen (probably approximate correct)
• Ähnlich zur VC-Theorie
• Berücksichtig rechnerische Komplexität
• Valient
Neyman-Pearson-Wald Entscheidungstheorie
• MinMax
• Bayes Optimal
Regularisierungstheorie
• Regularisierung: schlecht-gestellten Problemen->gut-gestellte Probleme
• Hadamard, Tikhonov
Prinzip der kleinsten Quadrate
• Gauss
• Entspricht Gauss'scher Likelihood
Algorithmische Statistik
• Fokus auf Vorhersageleistung (nicht Parameterschätzung)
• Breiman, Huber, Friedman
Wahrscheinlichkeitslehre
• Beispiel: Bester linearer Schätzer
• Eigentlich nicht Statistik, aber führt zu einfachen Termen (Korrelationen), die geschätzt werden können
MDL – Theorie (minimum decription lenght)
• Informationstheorie
• Rissanen, Wallace, Boulton
Information Bottleneck
• Tishby, Pereira, Bialek
Empirical Bayes (technicality)
• Type II likelihood
• Evidence Framework
(Subjektive) Bayes'sche Statistik:
• Auch subjektives Wissen kann in Form von Wahrscheinlichkeiten beschrieben werden und in die statistische Modellierung eingehen
• Beschreibend: ... Wie sich Menschen verhalten
• Normative (prescriptive): ...wie sich rationale Entscheidungssysteme (Menschen) entscheiden sollten
Objektive Bayes'sche Statistik
• Noninformative Priors (Jeffrey)
• Maximum Entropy Priors
• Grün: Frequent.
• Blau: Bayes
• Gold: Learn. Theory
• Rot: Rest

Lerntheorien

- Klassische Frequentistische Verfahren
  - C_p Statistik
  - Akaikes Informationskriterium (AIC)
- Bayes'sche Verfahren
  - Striktes Bayes: ich muss mich niemals entscheiden: Mitteln anstatt auswählen
  - Bayes'sche Modellauswahl, Bayesian Information Criterion (BIC)
- Moderne Frequentistische Verfahren
  - Minimum Description Length (MDL) Prinzip
  - Vapnik-Chervonenkis (VC) Theorie (Statistische Lerntheorie)
- **Wir werden evaluieren, wie diese Theorien die Differenz zwischen mittlerem Trainingsfehler und erwartetem Testfehler abschätzen!**

Klassische Frequentistische Verfahren

Frequentist Statistik (Fisher)

- Entwickelt primär von Fisher und ebenfalls durch Neyman und Pearson
- Die Theorie wurde vor dem weitverbreiteten Einsatz von Computern entwickelt; folglich lag die Betonung auf einfachen praktikablen Verfahren
- Eindeutig dominierender Ansatz; wird von der überwiegenden Anzahl von Statistikern verwendet
- Wesentlicher Vorteil: Einfachheit in der Handhabung; dadurch weite Verbreitung
- Der Ansatz benötigt keine a priori Verteilung und lässt die Daten für sich sprechen
- Basis des Ansatzes: Verhalten unter hypothetischer Wiederholung des Experimentes unter ähnlichen Umständen: *statistical procedures have to be assessed by their performance in hypothetical repetitions under identical conditions*

Frequentist Statistik (2)

- Vorgehensweise:
  - Wähle eine Statistik (d.h. eine sinnvolle Funktion der Daten, z.B. den Mittelwert)
  - Leite die Sampling Statistik ab (z.B.: wie verteilt sich der geschätzte Mittelwert um den wahren Mittelwert)
  - Messe die Plausibilität jedes möglichen Parametervektors
- Betonung liegt auf Parameterschätzung und nicht Prognosegüte
- Bei likelihoodbasierten Schätzern: das wahre Modell muss in der betrachteten Menge von Modellen sein
- Kritik (Bayesianer, Vapnik): Eine Menge von Tricks und ohne zugrundeliegende geschlossene Theorie; funktioniert nur bei Modellen mit wenigen Parametern, die sich gut schätzen lassen; die Theorie hat wenig zu sagen, wenn nur wenige Trainingsdaten vorliegen

- Frequentistische Ansätzte legen den Schwerpunkt auf Modellselektion und nicht so sehr auf Regularisierung

Frequentistische Ansätze

- Man vereinfacht das Problem, indem man den erwarteten Generalisierungsfehler nur an den Eingangspunkten des Trainingssatzes berechnet
- Daraus folgt, dass der abgeschätzte Generalisierungsfehler typischerweise zu klein ist
- Dieser *in-sample* Generalisierungsfehler lässt sich abschätzen zu

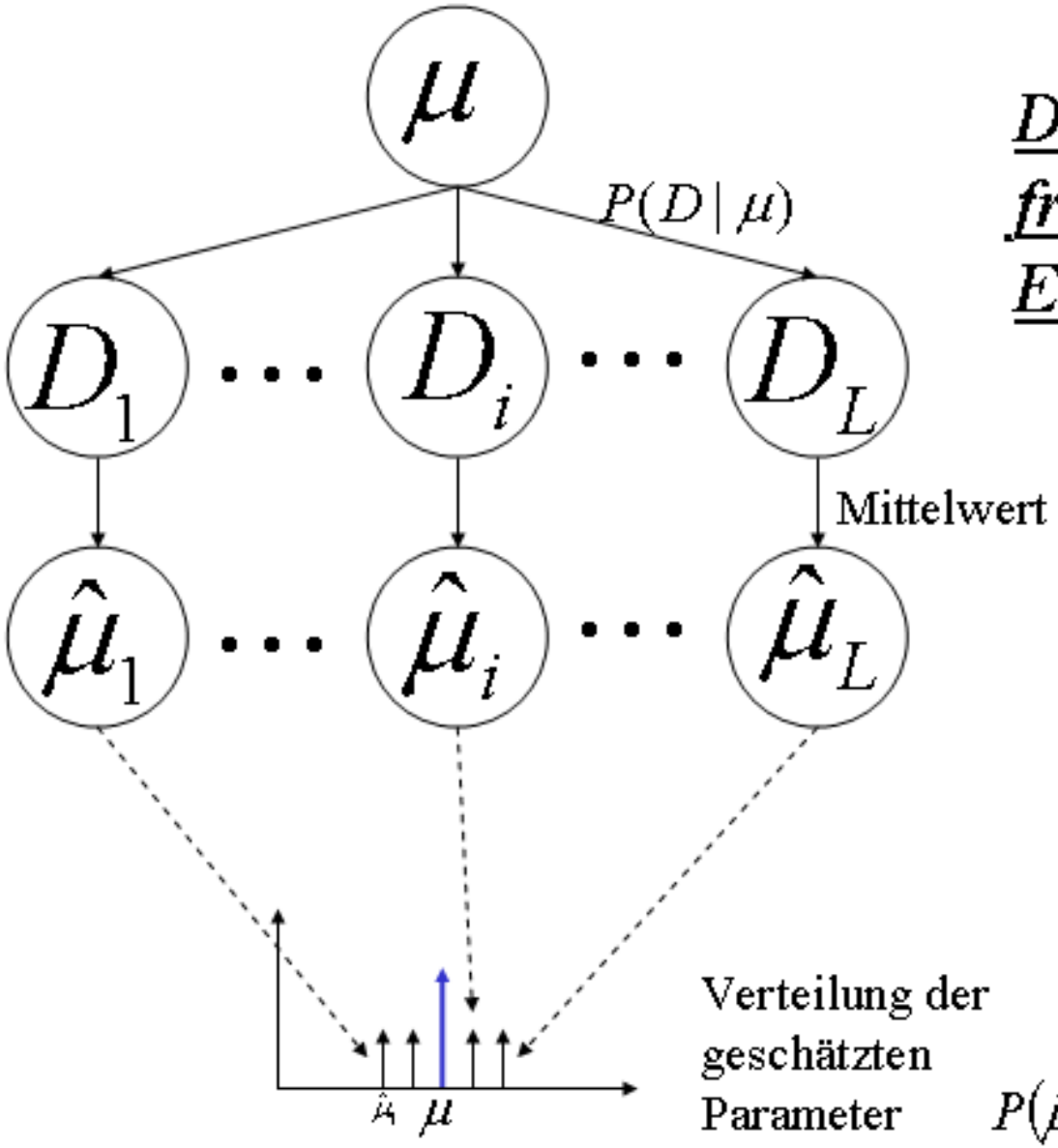
$$J_{in} = \frac{1}{N} \sum_{i=1}^{N} L[y_i, f(\mathbf{x}_i, \mathbf{w})] + Komplexitätsterm$$

Mallows' C_p Statistik

- Man erhält für Modelle, die linear in den Parameterm sind und quadratischem Fehler als Verlustfunktion

$$J_{in} = C_p = \frac{1}{N} \sum_{i=1}^{N} L[y_i, f(\mathbf{x}_i, \mathbf{w})] + 2\frac{M}{N}\hat{\sigma}^2$$

- M ist die Anzahl der Modellparameter
- Kann zur Auswahl der Eingänge benutzen (lineares Modell) oder zur Auswahl relevanter fester Basisfunktionen
- $\hat{\sigma}^2$ ist die geschätzte Rauschvarianz
- J_{in} ist der abgeschätzte Fehler an den Eingangsdaten der Trainingsdaten (*in-sample*) (und ist daher oft zu optimistisch)



Das frequentistische Experiment

Verteilung der geschätzten Parameter

$$P(\hat{\mu} \mid \mu) \propto \mathrm{N}\left(\mu, \sigma^2/N\right)$$

Beispiel: Schätzung des Mittelwertes

- Ansatz

$$x_i = \mu + \epsilon_i$$

  σ^2 sei die unbekannte Rauschvarianz

- Schätzung

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  Diese ist erwartungstreu und ergibt sich daraus, dass der Erwartungswert der Summe gleich der Summe der Erwartungswerte ist.

- Wie man leicht sieht ist die Varianz der Schätzung

$$var(\hat{\mu}) = \frac{\sigma^2}{N}$$

  Dies folgt daraus, dass für unabhängige Zufallsprozesse die Varianz der Summe gleich der Summe der Varianz der Elemente ist und weil $var(ax) = a^2 var(x)$, mit konstantem a.

- Die Varianz des geschätzten Parameters ist unkorreliert mit der Rauschvarianz der neuen Daten
- Deshalb ist der erwartete Testfehler die Summe aus Parametervarianz und Rauschvarianz :

$$J^{test} = \frac{\sigma^2}{N} + \sigma^2 = \frac{N+1}{N}\sigma^2$$

- Setzen wir eine erwartungstreue Schätzung der Rauscharianz ein

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 = \frac{N}{N-1}J^{train}$$

mit $J^{train} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$

- Damit ist

$$J^{test} = \frac{N+1}{N-1}J^{train}$$

$$= \frac{N-1+2}{N-1}J^{train} = J^{train} + 2\frac{1}{N-1}J^{train} = J^{train} + 2\frac{\widehat{\sigma}^2}{N}$$

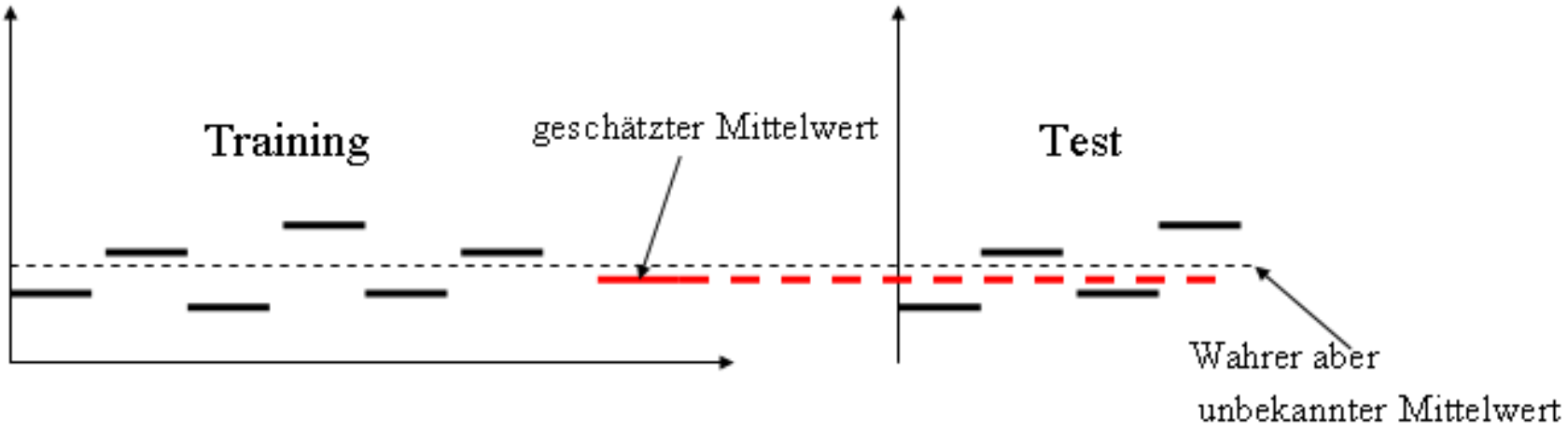
- Dieses Ergebnis kann generalisiert werden zu Modellen mit M Parametern und man erhält

$$J^{test} = \frac{N+M}{N-M} J^{train}$$

$$= J^{train} + 2\frac{M}{N}\hat{\sigma}^2$$

- Dies ist identisch zu C_P. Beachte, dass der Unterschied zwischen Trainingsfehler und erwartetem Testfehler proportional ist zur Anzahl freier Parameter und inverse proportional ist zur Anzahl der Trainingsdaten!

Testfehler für Schätzung eines Mittelwertes



- Varianz der Testdaten und der Trainingsdaten: σ^2
- Der empirische Mittelwert (rot) ist erwartungstreu mit Varianz $\frac{\sigma^2}{N}$

- Der erwartete mittlere Testfehler ist der erwartete quadratische Abstand zwischen geschätztem Mittelwert (rot) und den Testdaten und ist die Summe aus der Varianz der Testdaten und der Varianz der Schätzung:

$$\sigma^2 + \frac{\sigma^2}{N}$$

Akaikes Information Criterion (AIC)

- Man erhält für Modelle, bei denen die Log-Likelihood

$$l = \log L = \sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

optimiert wird Akaike's *Information Criterion* (minimiere:)

$$AIC = 2\left(-\frac{1}{N}\log L + \frac{M}{N}\right)$$

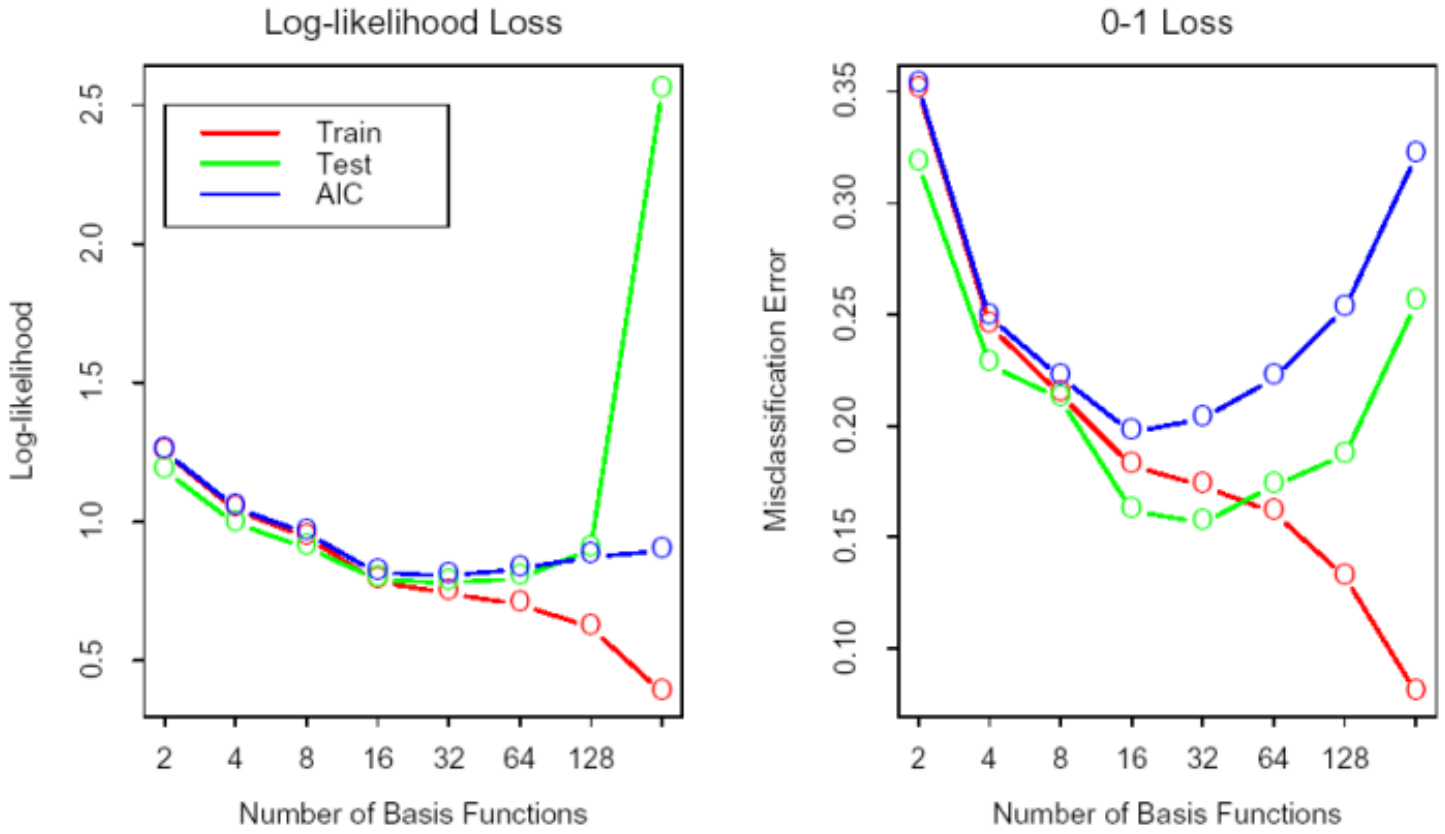
- $\frac{M}{N}$ ist eine Schätzung der Differenz zwischen mittlerer Trainings-Loglikelihood und mittlerer Test-Loglikelihood.
- AIC ist äquivalent zu C_p für Gauss Rauschen mit bekannter Rauschvarianz:

$$AIC = \frac{1}{\sigma^2}\frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i, \mathbf{w}))^2 + 2\frac{M}{N} = \frac{1}{\sigma^2}C_P$$

- $$-AIC/2 = \frac{1}{N} \log L - \frac{M}{N}$$

  schätzt die mittlere Loglikelihood von neuen Daten ab, deren Eingangswerte mit den Trainingsdaten übereinstimmen (*in-sample*)

AIC für Likelihood Kostenfunktion und für 1/0 Kostenfunktion



Misspezifizierte Modelle

- Der Likelihood-Ansatz (und ebenso der Bayes'sche Ansatz) nimmt an, dass sich das wahre Modell in der Klasse der betrachteten befindet
- Man kann jedoch zeigen, dass im Fall der Misspezifikation der ML-Ansatz definierte und sinnvolle Ergebnisse liefert
- Betrachten wir als Abstand zwischen wahrer Verteilung $P(\mathbf{x})$ und approximitiver Verteilung $P_\theta(\mathbf{x})$ mit Parametern θ den sogenannten Kullback-Leibler Abstand (KL-Divergenz)

$$KL(P\|P_\theta) = \int P(\mathbf{x}) \log \frac{P(\mathbf{x})}{P_\theta(\mathbf{x})} d\mathbf{x}$$

- Der Kullback-Leibler Abstand ist gleich Null, wenn beide Verteilungen gleich sind und ist ansonsten größer Null. Beachte, dass der KL-Abstand unsymmetrisch ist: $KL(P\|P_\theta) \neq KL(P_\theta\|P)$

- Approximiert man die wahre unbekannte Verteilung durch die Samples, erhält man die negative Loglikelihood

$$KL(P\|P_\theta) \approx -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta(\mathbf{x}_i)$$

- Man kann nun zeigen, dass unter schwachen Regularitätsbedingungen ein Modell, welches die Loglikelihood maximiert asymptotisch zu Parametern konvergiert, so dass der Abstand zwischen wahren und approximativem Modell im Sinne der KL-Divergenz minimal ist

- Dies bedeutet, dass auch wenn die wahre Verteilung nicht in der Klasse der betrachteten Modelle ist, der ML-Ansatz sinnvolle Ergebnisse liefert!

Varianten der frequentistischen Statistik

- Fiducial (Bezugs) Inference, Pivotal Inference, Structural Inference: Versuche, Parameterwahrscheinlichkeiten abzuleiten, ohne die Bayes'sche Theorie anzuwenden
- Regularisierung: Stein Schätzer, Ridge Regression, ...
- Robuste Statistik (Huber)

Varianten der frequentistischen Statistik: Algorithmische Ansätze

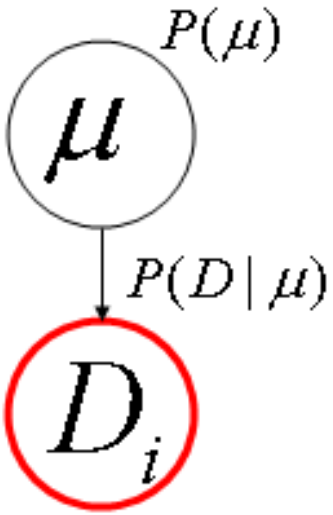
- Breiman, Hastie, Friedman, Huber, ...
- Nicht Suche nach den besten Parameterm steht im Vordergrund, sonder nach dem System, welches die besten Prognosen liefert
- Techniken: Kreuzvalidierung
- Analyse eher frequentistisch (Bias - Variance)
- Pragmatischer Ansatz
- Wahres Modell ist nicht der Fokus der Analyse

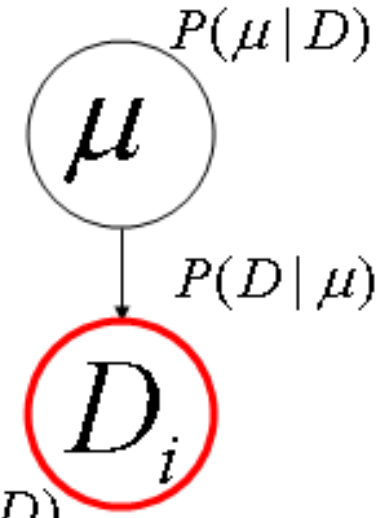
Bayes'sche Ansätze

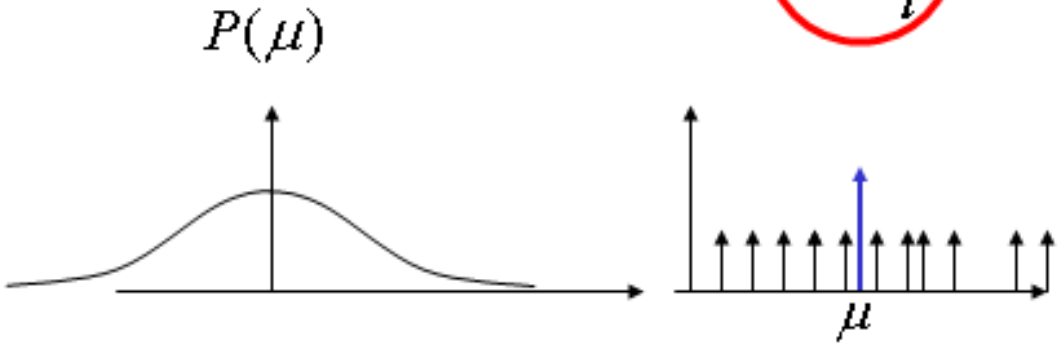
Bayes'scher Ansatz

- Axiomatisches Fundament: Entscheidungen unter Unsicherheit
- Wissenschaftlich komplettes Modell
- Man sollte ehrlich seine Annahmen explizit machen, und basierend auf diesen dann Entscheidungen treffen
- Parameter und Variable werden gleichberechtigt als Zufallsvariable behandelt; streng genommen gibt es nur das Problem der Inferenz (und nicht der Parameterschätzung)
- Prognose im Vordergrund, nicht Parameterschätzung
- Bayes'sche Ansätze neigen wesentlich weniger zum Überanpassen
- Nachteile
  - Inferenz führt zu komplexen Integralen, die numerisch approximiert werden (MCMC, Markov Chain Monte Carlo)

- Die aufwendige Maschinerie bringt einen vom eigentlichen Problem weg (explorative Analyse)
- Das wahre Modell muss im Satz der betrachteten Modelle sein (Vapnik: einziges aber schwerwiegendes Problem)







$$P(\mu) \propto \mathrm{N}(0, \alpha^2)$$

$$P(\mu \mid D) \propto \mathrm{N}\left(\frac{mean}{1 + \frac{\sigma^2}{N\alpha^2}}, \frac{\sigma^2}{N + \sigma^2 / \alpha^2} \right)$$

Das Bayes'sche

Experiment

Bayes'scher Ansatz: Varianten

- Subjektiver Bayes: Konsequente Einziehung von Vorwissen
- Objektiver Bayes: Definition von a priori Verteilungen, so dass die a prior Annahme möglichst wenig Einfluss auf das Ergebnis hat
  - Uninformative Prior (Jeffrey)
  - Maximum Entropie Ansatz
- Empirial Bayes: Schätzung von Hyperparametern
  - Evidence Framework (Type II Likelihood): Modellauswahl nach $P(D|\mathcal{M})$

Bayes'sche Modellauswahl

- Wenn ich denn doch ein Modell auswählen muss ...
- A posteriori Modellwahrscheinlichkeit

$$P(\mathcal{M}|D) \propto P(\mathcal{M})P(D|\mathcal{M})$$

- Typischerweise nimmt man an, dass alle Modelle gleich-wahrscheinlich sind (a priori)
- Somit ist der entscheidende Term (marginal likelihood, evidence)

$$P(D|\mathcal{M}) = \int P(w|\mathcal{M})P(D|w)dw$$

Laplace Approximation der Marginal Likelihood

- log $P(D|\mathcal{M})$ wird asymptotisch gaussförmig, allerdings ist das Integral nicht zu Eins normiert;
- Man behält nun nur die Terme, die von N abhängen. Dann erhält man

$$\log P(D|\mathcal{M}) \approx \log P(D|\hat{\mathbf{w}}_{MAP}, \mathcal{M}) - \frac{M}{2} \log N$$

- Übungsaufgabe: leiten Sie diese Approximation her

Bayesian Information Criterion (BIC)

- BIC ist 2 Mal diesem Ausdruck (man ersetzt die MAP Parameterschätzung durch die ML-Parameterschätzung) (minimiere)

$$BIC = -2 \log L + M \log N$$

und die mittlere vorhergesagte Loglikelihood

$$-\frac{1}{2N} BIC = \frac{1}{N} \log L - \frac{M}{N} \frac{1}{2} \log N$$

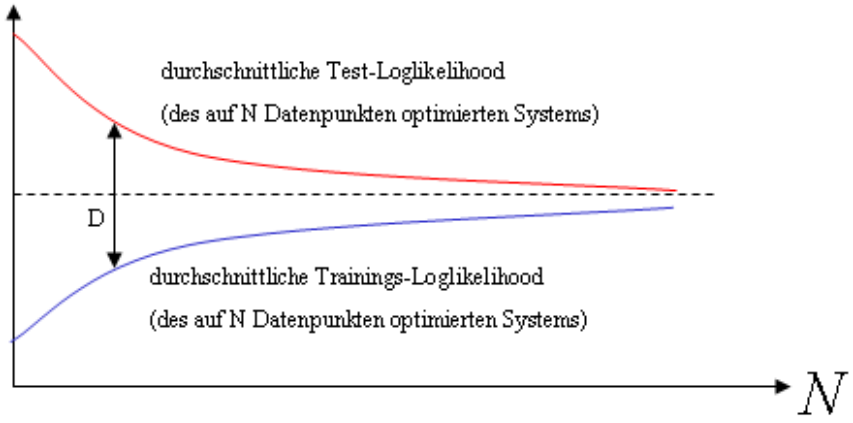
Vergleiche

$$-AIC/2 = \frac{1}{N} \log L - \frac{M}{N}$$

- $\frac{M}{N} \frac{1}{2} \log N$ ist eine Schätzung der Differenz zwischen mittlerer Trainings-Loglikelihood und mittlerer Test-Loglikelihood.

- Die BIC Korrektur ist um den Faktor $\frac{1}{2} \log N$ größer und verringert sich langsamer mit $(\log N)/N$ mit der Anzahl der Trainigsdaten

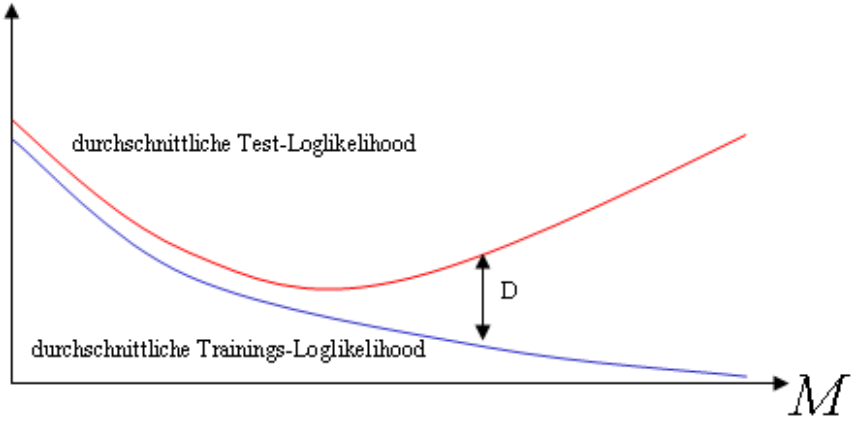
Vergleich: AIC und BIC



Schätzung von D:

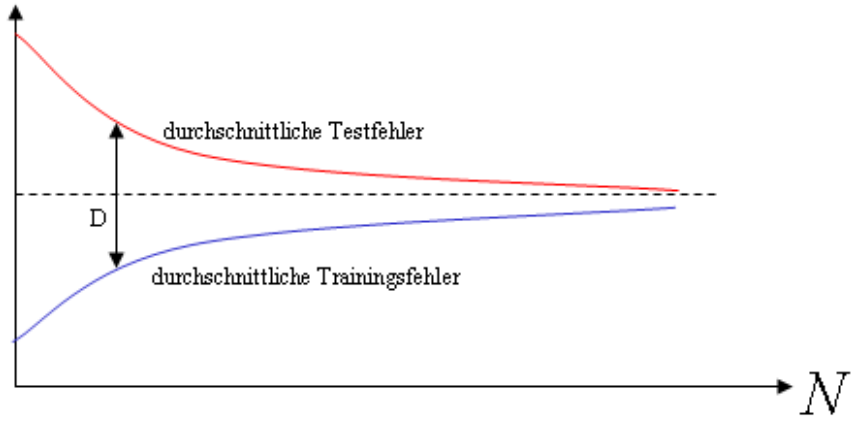
AIC: $\hat{D} = \frac{M}{N}$

BIC: $\hat{D} = \frac{M}{N}\frac{1}{2}\log N$



Mit einer zunehmenden Anzahl von Datenpunkten N verringert sich D (d.h. die Überanpassung), mit zunehmender Komplexität M erhöht sich D

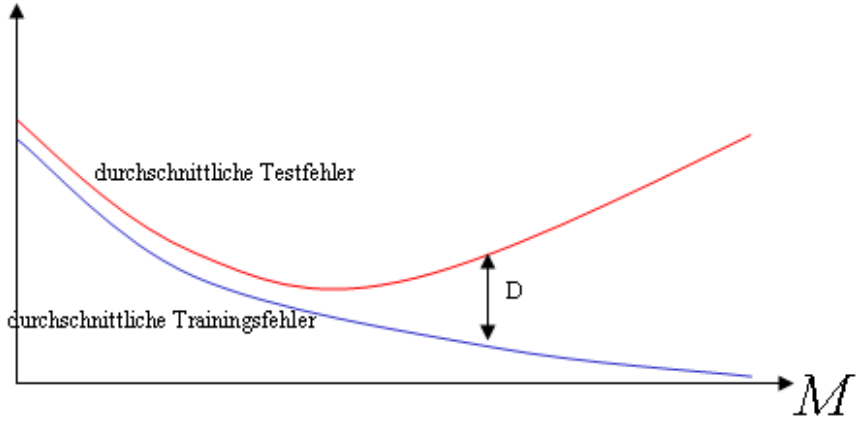
AIC und BIC: Spezialfall Gauß'sche Likelihood, quadratischer Fehler



Schätzung von D:

AIC: $\hat{D} = \frac{M}{N}\sigma^2$

BIC: $\hat{D} = \frac{M}{N}\frac{\sigma^2}{2}\log N$



Mit einer zunehmenden Anzahl von Datenpunkten N verringert sich D (d.h. die Überanpassung), mit zunehmender Komplexität M erhöht sich D

Moderne Frequentistische Verfahren

Minimum Description Length

- Basierend auf dem Konzept der algorithmischen Komplexität (Kolmogorov, Solomonoff, Chaitin)
- Auf Basis dieser Ideen: Rissanen (und Wallace, Boulton) führten das Prinzip der minimum description length (MDL) ein
- Unter einigen Vereinfachungen wird das MDL Kriterion identisch zum BIC Kriterion (siehe Appendix)

Statistical Learning Theory

- Start: Kolmogorov, Glivenko, Cantelli
- Der Vater der SLT: Vladimir Vapnik
- Ziel: Gegeben eine Menge von Funktionen, die nicht die wahre Funktion enthalten müssen, wähle die optimale Funktion aus
- Forderung der Konsistenz: Asymptotisch soll die beste Funktion ausgewählt werden
- Im Zentrum steht hier wieder die Differenz zwischen Trainingsfehler $R_{emp}(f)$ und Testfehler $R(f)$. Im Gegensatz zu vorher steht hier jedoch nicht die Differenz zwischen Trainingsfehler und *erwartetem* Testfehler im Vordergrund, sondern die Theorie fokussiert auf die Berechnung *einer oberen* Schranke zwischen Trainingsfehler und Testfehler!

STL (2)

- Worst Case Analysis (MinMax) (one-sided uniform convergence)

$$\lim_{N\to\infty} P\left(\max_{f\in F} |R(f) - R_{emp}(f)| > \epsilon\right) = 0, \forall \epsilon > 0$$

  (die gilt **für alle** $f : A \leq R(f) \leq B$ mit beliebigen Schranken A, B)

- Vapnik argumentiert, dass nur eine Worst-Case-Analyse zu konsistenten nicht-trivialen Resultaten führt

- Nachteile: die berechneten Schranken sind in der Praxis zu konservativ und entsprechen nicht dem tatsächlichen Generalisierungsfehler

Vapnik-Chervonenkis (VC-) Theorie (Statistical Learning Theory)

- Die VC-Theorie ist verteilungsfrei, das heißt sie macht keine Annahmen über eine zugrundeliegende Verteilung; speziell muss sie auch nicht annehmen, dass die wahre Verteilung sich in der Klasse der betrachteten Verteilungen befindet
- Die einzige wesentliche Annahme: Daten werden von einer *festen* Verteilung $P(\mathbf{x})$ generiert
- Zielgrößen werden von $h(\mathbf{x})$ generiert (im einfachsten Fall und hier ohne Rauschen und binär)
- Man versuche $h(\mathbf{x})$ mit $f(\mathbf{x})$ zu approximieren. $f(x)$ sei ein Mitglied einer Klasse von Funktionen $F(\mathbf{x})$.

VC-Theorie (2)

- Die mittlere Generalisierungsperformanz (Risiko) ist

$$R(f) = \int P(\mathbf{x}) l(h(\mathbf{x}), f(\mathbf{x})) d\mathbf{x}$$

  wobei $l(a, b) = 0$, falls $a = b$ ist und 1 anderenfalls

- Sei

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} l(h(\mathbf{x}_i), f(\mathbf{x}_i))$$

  das empirische Risiko

VC-Theorie (3)

- Wie ist im schlimmsten Fall, d.h. für die ungünstigste Funktion $f \in F$ und für die ungünstigste Verteilung der Trainingsdaten der Unterschied zwischen $R(f)$ und $R_{emp}(f)$ für N Datenpunkte? D.h. wie ist,:

$$\max_f |R(f) - R_{emp}(f)|$$

- Vapnik hat gezeigt, dass unabhängig vom speziellen h und den speziellen Trainingsdaten gilt:

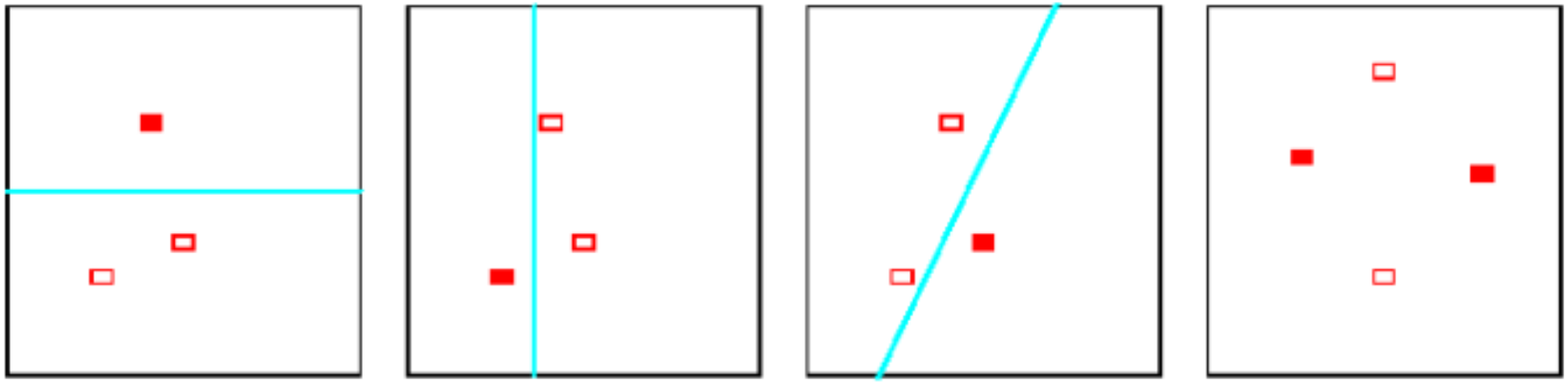
$$P(\max_{f \in F} |R(f) - R_{emp}(f)| > \epsilon) \leq \text{bound} = 4\Delta(2N) \exp(-\epsilon^2 N/8)$$

VC-Dimension

$$P(\max_{f \in F} |R(f) - R_{emp}(f)| > \epsilon) \leq 4\Delta(2N) \exp(-\epsilon^2 N/8)$$

- $\Delta(N)$ (*growth function*) ist eine obere Schranke für die maximale Anzahl der verschiedenen binären Funktionen, die $F(\mathbf{x})$ auf (mindestens einer Menge von) N Daten implementieren kann
- $\Delta(N)$ wächst entweder asymptotisch wie 2^N für alle N oder ist nach oben begrenzt durch $N^{d_{VC}} + 1$, wobei d_{VC} die (berühmte) VC-Dimension von $F(\mathbf{x})$ ist; im ersten Fall ist d_{VC} unendlich und das Lernsystem erlaubt keine Generalisierung (aus Hertz, Krogh, Palmer: Introduction to the theory of neural computation)
- Die VC-Dimension einer Funktionenklasse F ist die größte Anzahl von Datenpunkten (in mindestens einer Anordnung), die von Mitgliedern von F ge-*shattered* werden können
- Für einen linearen Klassifikator ist $d_{VC} = M$, d.h. gleich der Anzahl der freien Parameter (Anzahl der Eingangsvariablen plus 1)

- Shattered: Egal wie ich Zielwerte den Datenpunkten zuordne, ein Mitglied der Klasse kann es korrekt Klassifizieren (für mindestens eine Anordnung der Eingangsvektoren)

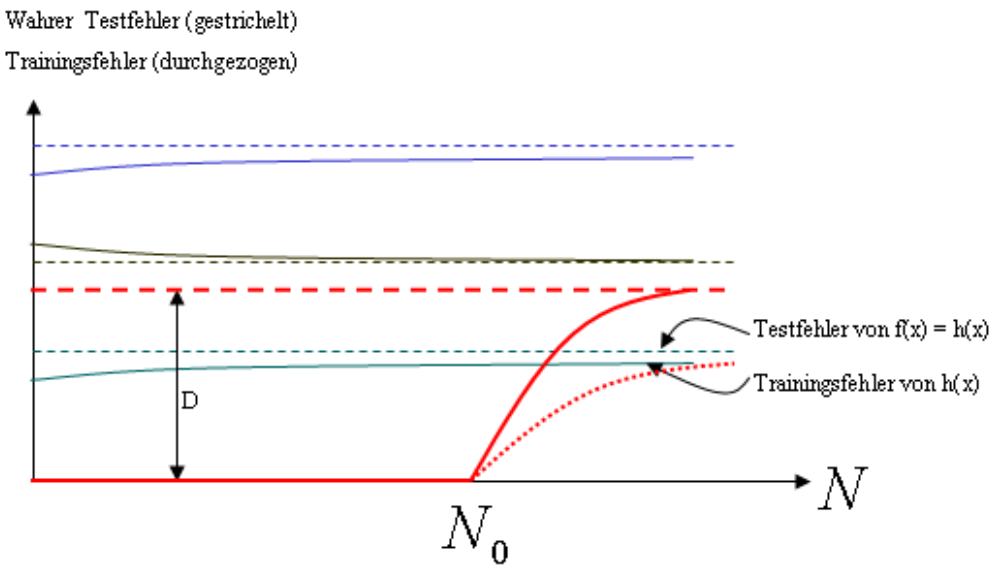


- Prinzip der *Structural Risk Minimization* (SRM): wähle die *Modellklasse*, für welches

$$R_{emp}(f) + \text{bound}$$

minimal ist

VC-Theorie



Die Abb. zeigt 4 verschiedene f für einen bestimmten Trainingsdatensatz. Die VC-Theorie liefert eine Aussage über eine Schranke für D. In dem Beispiel ist entsprich die rote Kurve einer f, die bis zur Datenmenge N_0 alle Daten perfekt anpasst. Wenn es mindestens eine Anordnung der Eingangsdaten gibt, so dass für eine beliebige Anordnung der Zielwerte, es mindestens ein f gibt, für das N_0 unendlich ist, dann ist die VC-Dimenion der Funktionsklasse unendlich und es ist, im Sinne von VC, keine Generalisierung möglich. Rot gepunktet entspricht dem f, welches für ein N den Trainingsfehler minimiert. Beachte, dass dies für jedes N in der Regel ein anderes f ist.

Hier ist ein subtiler Punkt: der Trainingsfehler ist ein unverzerrter Schätzer für jedes beliebiges f. Für ein endliches N wird es aber ein f geben, wo der Trainingsfehler durch Zufall noch sehr viel geringer ist als der Testfehler. Dies ist natürlich das f, welches den Trainingsfehler minimiert! Dies bedeutet. Wenn ich jeweils das f mit dem besten Trainingsfehler auswähle, habe ich jeweils eine verzerrte Aussage.

Vapnik-Chervonenkis (VC-) Theorie: Vorteile und Nachteile

- Vorteil: es muss nur angenommen werden, dass $P(\mathbf{x})$ fest ist; weder eine prior Verteilung noch eine Likelihood Funktion muss definiert werden
- Nachteile: Die VC-Dimension lässt sich für viele interessante Klassen von Funktionen nicht berechnen; nur weniger gute oder schlechte Grenzen sind verfügbar
- Als worst-case Theorie ist die Übertragbarkeit auf den *average case* nur begrenzt möglich

APPENDIX

VC und Supportvektormaschine

- Eine Funktionsklasse F_A ist definiert durch alle linearen Klassifikatoren mit $\sum_{i=1}^{M-1} w_i^2 \leq A^2$ implying that $\mathcal{C} \leq 1/A$. In diesem Fall kann die VC-Dimension kleiner als M sein; es gilt: je größer der Margin $\mathcal{C}$ umso kleiner die VC-Dimension
- Dies bedeutet, dass gegeben N das Modell ausgewählt wird, welches mit maximalem Margin die Kostenfunktion minimiert (da diese auch die Funktion ist wo der Abstand zwischen Train und Test in der Regel maximal ist in der betrachteten Funktionsklasse minimal ist)
- Man berechnet den bound und wählt die Funktionsklasse (mit dem spezifischen A), für welches die Summe aus bound und Fehler minimal ist
- Da keine guten bounds existieren verwendet man Cross-Validierung zur Einstellung von A

MDL: Modellannahmen

- Eine (typische) Codelänge für ein typisches Muster y in einem optimalen Code ist $-\log_2 P(\mathbf{y})$ (Shannon)
- Wir wollen die Zielwerte der Trainingsdaten $\{y_i\}_{i=1}^N$ übertragen
- Naiver Ansatz: wir übertragen die Daten, die eine mittlere Codelänge $-\log_2 P(y)$ besitzen
- Modellansatz:
  - Sender und Empfänger kennen beide die Eingangsdaten und die priori Verteilung und die funktionelle Form der Likelihood; Ziel ist die effizienteste Übertragung der Daten $\mathbf{y}$.
  - Wir trainieren ein Modell und erhalten den Parametervektor $\widehat{\mathbf{w}}$
  - Wir übertragen zunächst $\widehat{\mathbf{w}}$ mit erwarteter Codelänge $-\log_2 P(\widehat{\mathbf{w}})$ und dann die Daten mit erwarteter Codelänge $-P(\mathbf{y}|\widehat{\mathbf{w}})$

– Die gesamte erwartete Codelänge (description length) ist somit

$$-\log P(\hat{\mathbf{w}}) - \log P(D|\hat{\mathbf{w}})$$

welche typischerweise geringe ist als $-\log_2 P(y)$

- Nach dem MDL (minimum descriprion length Modell) Prinzip ist das Modell optimal, für welches MDL minimal ist
- Die DL kann angenähert werden zu (siehe Appendix)

$$E(DL) \approx -\log L(\hat{w}) - \log P(\hat{w}) \approx -\log L(\hat{w}) + \frac{M}{2} \log N$$

- Hier wird Rissanen's MDL Kriterium equivalent zur Bayes'schen Modellauswahl, d.h. approximativ zu BIC.
- MDL hat eine längere Entwicklung hinter sich, die diese kurze Diskussion nur unzureichend wiederspiegelt. Für eine weitergehende Diskussion: www.gruenwald.nl: A tutorial introduction to the MDL principle.

MDL: Bezug zur Informationstheorie

- Ziel ist die (wiederholte) Übertragung der Werte einer Zufallsvariablen $\mathbf{x}$ mit Verteilung $P(\mathbf{x})$
- Shannon's Theorem (Source Coding Theorem) sagt aus, dass die mittlere Codelänge (*description length*, DL) eines Codes größer oder gleich der Entropie ist

$$E(DL) \geq -\sum_{\mathbf{x}} P(\mathbf{x}) \log_2 P(\mathbf{x})$$

  $DL =$ Länge des binären Codes

- Ein optimaler Code würde die Gleichheit erfüllen (Shannon Limit) und würde dem Wert $\mathbf{x}$ die Länge $-\log_2 P(\mathbf{x})$ zuordnen
- Dies bedeutet, dass häufigere Muster einen kürzeren Code erhalten sollten
- Eine (typische) Codelänge für ein typisches Muster $\mathbf{x}$ ist $-\log_2 P(\mathbf{x})$

MDL: Modellannahmen

- Wir wollen die Zielwerte der Trainingsdaten $\{y_i\}_{i=1}^N$ übertragen
- Sender und Empfänger kennen beide die Eingangsdaten und die funktionelle Form von a priori Verteilung und Likelihood; Ziel ist die effizienteste Übertragung der Daten $\mathbf{y}$.
- Wir übertragen erst den Parametervektor $\mathbf{w}$ mit $P(\mathbf{w})$
- ... und dann die Ausgänge mit $P(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathcal{M})$
- Wir gewinnen, da $P(\mathbf{y})$ ohne Regressionsmodell eine sehr viel kleinere Wahrscheinlichkeitsdichte besitzt wie mit Regressionsmodell und $P(\mathbf{y}|\mathbf{w}, \mathbf{X}, \mathcal{M})$

Rissanen's Minimum Description Length (Modellselektion)

- Betrachten wir nun ein Modell $\mathcal{M}$ mit a priori Parameter Verteilungen $P(\mathbf{w})$ und Likelihoods $P(D|\mathbf{w})$
- Angenommen, dass der Parameter Schätzer $\hat{\mathbf{w}}$ und die Likelihood $P(D|\hat{\mathbf{w}})$ typischen Werten entsprechen, so ist die typische Codelänge gleich

$$-\log P(\hat{\mathbf{w}}) - \log P(D|\hat{\mathbf{w}})$$

  Dies bedeutet, dass man für die effizienteste Übertragung das Modell wählen sollte, für das diese Summe minimal ist

MDL und BIC

- Eine genauere Analyse berücksichtigt, dass eine ungenaue Kodierung von $\mathbf{w}$ äquivalent zu zusätzlichem Rauschen auf der Zielgröße ist
- Man kann argumentieren, dass der Parametervektor $\mathbf{w}$ in jeder Dimension nur mit $\sqrt{N}$ Bins pro Dimension übertragen werden muss. Dies bedeutet, dass bei mehr Daten man mit einer besseren Kodierung der Parameter gewinnt. Unter der Annahme von Uniformität ist der Komplexitätsterm

$$\log P(\mathbf{w}) \rightarrow \log(1/\sqrt{N})^M = -\frac{M}{2} \log N$$

und MDL ist äquivalent zu BIC.