

Vorlesung Wissensentdeckung in Datenbanken

Bias und Varianz

Kristian Kersting, (Katharina Morik), Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

22.05.2014

Gliederung

- 1 Wiederholung: Lineare Modelle zur Klassifikation und Regression
 - Klassifikation und Regression
 - Die schon bekannten Lineare Modelle
 - Geometrie linearer Modelle: Hyperebenen

- 2 Bias-Varianz
 - Exkurs: Erwartungswert
 - Bias und Varianz bei linearen Modellen

Disclaimer

Die Wiederholung der linearen Funktionen dient der Vorbereitung und Festlegung der Nomenklatur/Notation auf den Bias/Variance Tradeoff und auf die Stützvektormaschinen.

Notation

Sei $X = \{X_1, \dots, X_p\}$ eine Menge von p vielen Zufallsvariablen und $Y \neq \emptyset$ eine Menge. .

Ein **Beispiel** (oder *Beobachtung*) $\vec{x}$ ist ein konkreter p -dimensionaler Vektor über diese Zufallsvariablen.

Eine **Menge von N Beispielen** $\mathbf{X} = \{\vec{x}_1, \dots, \vec{x}_N\}$ können wir dann als $(N \times p)$ -Matrix auffassen:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,p} \end{pmatrix}$$

Dabei entspricht jede **Zeile** $\vec{x}_i$ der **(Daten-)Matrix** $\mathbf{X}$ einem **Beispiel**.

Überwachtes Lernen: Klassifikation und Regression

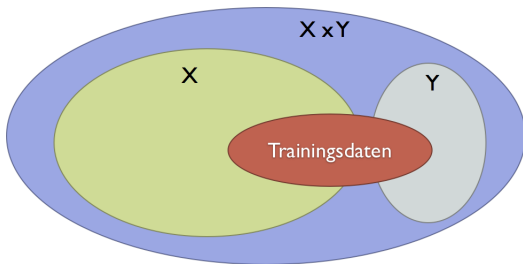
Beim **überwachten Lernen** (darum geht es hier), ist zusätzlich zu jeder Beobachtung $\vec{x}$ ein *Label (Klasse)* y gegeben, d.h. wir haben Beobachtungen $(\vec{x}, y) \in X \times Y$.

Für $Y = \mathbb{R}$ versuchen für unbekanntes $\vec{x}$ den Wert y vorherzusagen:
Regression.

Für die **Klassifikation** ist Y eine diskrete Menge.

Überwachtes Lernen (Klassifikation/Regression) auf Trainingsdaten

Wovon gehen wir also bei der Klassifikation und der Regression aus?
Was ist unser Ziel?



- Wir suchen *die wahre Funktion* $f : X \rightarrow Y$ mit

$$f(\vec{x}) = y \quad \forall (\vec{x}, y) \in X \times Y$$

- Wir haben jedoch nur eine Teilmenge der Beobachtungen gegeben (Trainingsdaten)

Klassifikation und Regression

Auf Grundlage der Trainingsdaten suchen wir eine möglichst gute Annäherung $\hat{f}$ an die *wahre Funktion* f .

Die Funktion $\hat{f}$ bezeichnen wir auch als das gelernte **Modell**.

Haben wir ein Modell $\hat{f}$ gelernt, so liefert uns dieses Modell mit

$$\hat{y} = \hat{f}(\vec{x})$$

für *neue Daten* $\vec{x} \in X$ eine Vorhersage $\hat{y} \in Y$.

Klassifikation und Regression

Im Falle der **Regression** läßt sich so für zuvor unbekannte $\vec{x} \in X$ der Wert

$$\hat{y} = \hat{f}(\vec{x})$$

mit $\hat{y} \in \mathbb{R}$ vorhersagen.

Ein Regressionsmodell $\hat{f}$ läßt sich auch für die **Klassifikation** nutzen, bei der z.B. $\hat{y} \in \{-1, +1\}$ vorhergesagt werden sollen:

$$\hat{y} = \begin{cases} +1, & \text{falls } \hat{f}(\vec{x}) \geq \theta \\ -1, & \text{sonst} \end{cases}$$

Hier ist θ ein vorgegebener Schwellwert.

Beispiel: Klassifikation

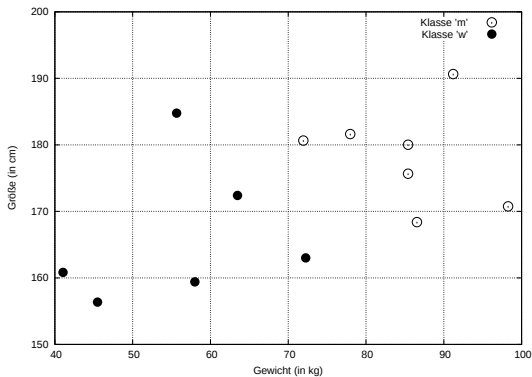
Gegeben seien Gewicht (X_1) und Größe (X_2) einiger Personen (ID) und ein Klassen-Label $y \in \{m, w\}$:

ID	X_1	X_2	Y
x_1	91	190	m
x_2	60	170	w
x_3	41	160	w
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Es wird nun eine Funktion $\hat{f}$ gesucht, die für neue Daten $\vec{x}$ das Attribut Y (Geschlecht) voraussagt, also

$$\hat{y} = \begin{cases} m, & \text{falls } \hat{f}(x) > \theta \\ w, & \text{sonst} \end{cases}$$

Beispiel: Klassifikation



Lineare Modelle

Welche Art von Funktionen sind denkbar?

Sie kennen ja schon **Lineare Funktionen** als einfachste Funktionenklasse:

$$y = f(x) = mx + b \quad \text{Gerade im } \mathbb{R}^2$$

Allerdings betrachten wir als Beispielraum den $\mathbb{R}^p$, d.h. wir brauchen eine verallgemeinerte Form:

$$y = f(\vec{x}) = \sum_{i=1}^p \beta_i x_i + \beta_0 \quad \text{mit } \beta_0 \in \mathbb{R}, \vec{x}, \vec{\beta} \in \mathbb{R}^p \quad (1)$$

Die Funktion f wird also durch $\vec{\beta}$ und β_0 festgelegt und sagt uns für ein gegebenes $\vec{x}$ das entsprechende y voraus

Notation, Vereinbarungen

Bei genauerer Betrachtung von Formel (1) läßt sich $\sum_{i=1}^p \beta_i x_i$ als Matrizenmultiplikation oder Skalarprodukt schreiben, also

$$y = \sum_{i=1}^p \beta_i x_i + \beta_0 = \vec{x}^T \vec{\beta} + \beta_0 = \langle \vec{x}, \vec{\beta} \rangle + \beta_0$$

Zur einfacheren Darstellung von f , wird β_0 in den Vektor $\vec{\beta}$ codiert, indem jedes Beispiel $x = (x_1, \dots, x_p)$ aufgefasst wird als $(p+1)$ -dimensionaler Vektor

$$(x_1, \dots, x_p) \mapsto (1, x_1, \dots, x_p)$$

Dies ermöglicht die Darstellung von f als:

$$y = f(\vec{x}) = \sum_{i=0}^p \beta_i x_i = \vec{x}^T \vec{\beta} = \langle \vec{x}, \vec{\beta} \rangle$$

Zur Erinnerung: Skalarprodukt

Das Skalarprodukt ist
definiert durch

$$\langle \vec{v}, \vec{w} \rangle = \vec{v}^T \vec{w} = \sum_{i=1}^p v_i w_i$$

Beispiel:

	$\vec{w} :$	
	4	
	5	
	6	
$\vec{v}^T : 1 \quad 2 \quad 3$		$1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 32$

Was haben wir nun gemacht?

Wir haben (bei der Beschränkung auf lineare Modelle) nun eine Darstellung für das, was wir *lernen* wollen:

$$y = \hat{f}(\vec{x}) = \vec{x}^T \vec{\beta} = \langle \vec{x}, \vec{\beta} \rangle$$

Wir haben die Zielfunktion $\hat{f}$ in Abhängigkeit von $\vec{\beta}$ geschrieben und müssen *nur noch* das passende $\vec{\beta}$ finden.

Veranschaulichung was das bedeutet

Unser Beispiel betrachtet ja eine binäre Klassifikation ($Y = \{-1, +1\}$, das Geschlecht):

- Was passiert dabei eigentlich anschaulich (geometrisch)?
- Wie klassifiziert unser $\hat{f}$ die Daten?
- Wie wirkt sich die Wahl von $\vec{\beta}$ aus?

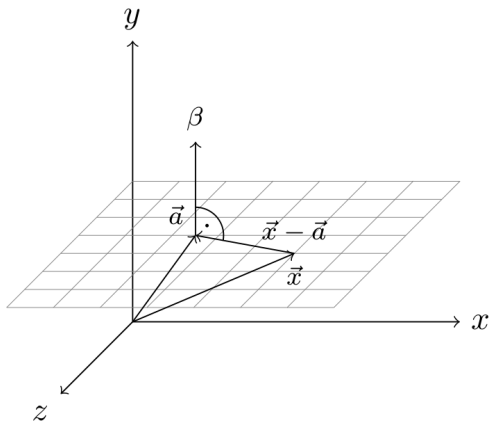
Zur Erinnerung: Ebenengleichung

Sei $V = \mathbb{R}^p$ ein Vektorraum, dann ist eine **Hyperebene** H ein $(p - 1)$ -dimensionaler affiner Untervektorraum.

H lässt sich über einen **Stützvektor** $\vec{a}$ und einen **Normalenvektor** $\vec{\beta}$ mit der **Ebenengleichung** schreiben

$$H = \left\{ x \in \mathbb{R}^p \mid \langle \vec{\beta}, \vec{x} - \vec{a} \rangle = 0 \right\}$$

Beispiel Ebene, Stützvektor, Normalenvektor



(Hyper-) Ebene im $\mathbb{R}^3$ mit Normalenvektor $\vec{\beta}$ und Stützvektor $\vec{a}$. Falls $\langle \vec{\beta}, \vec{x} - \vec{a} \rangle = 0$, also $\vec{\beta}$ und $\vec{x} - \vec{a}$ orthogonal zueinander sind, befindet sich $\vec{x}$ auf der Ebene.

Hesse Normalform

Multiplizieren wir die Ebenengleichung aus und setzen $\beta_0 = \langle \vec{\beta}, \vec{a} \rangle$,
dann ist

$$\langle \vec{\beta}, \vec{x} \rangle - \beta_0 = 0$$

in **Hesse Normalform**, falls $||\vec{\beta}|| = 1$.

Zur Erinnerung: Skalarprodukt

Das Skalarprodukt ist
definiert durch

$$\langle \vec{v}, \vec{w} \rangle = \vec{v}^T \vec{w} = \sum_{i=1}^p v_i w_i$$

aber auch durch den
Kosinus mit

$$\langle \vec{v}, \vec{w} \rangle = \|\vec{v}\| \cdot \|\vec{w}\| \cdot \cos(\angle(\vec{v}, \vec{w}))$$

Beispiel:

				$\vec{w} :$
				4
				5
				6
$\vec{v}^T :$	1	2	3	$1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 32$

Zur Erinnerung: Euklidische Länge

Euklidische Länge oder Norm

$$\|\vec{\beta}\| = \sqrt{\sum_{i=1}^p \beta_i^2} = \sqrt{\vec{\beta}^T \vec{\beta}} = \sqrt{\langle \vec{\beta}, \vec{\beta} \rangle}$$

weil $\|\vec{\beta}\|^2 = x_1^2 + \dots x_p^2$ (Pythagoras)

Beispiel: $\vec{\beta} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$

$$\|\vec{\beta}\| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

Normiert heißt ein Vektor, wenn er die (Euklidische) Länge 1 hat.

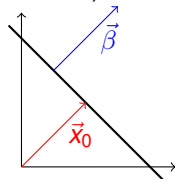
Abstandsberechnung durch Hesse Normalform

Sei $\vec{x}_0$ der Vektor, dessen Länge der Abstand vom Ursprung zur Ebene in **Hesse Normalform** ist. Dieser muss orthogonal zur Ebene liegen und somit parallel zu $\vec{\beta}$. Seien nun $\vec{\beta}$ und $\vec{x}_0$ gleichgerichtet, dann gilt

$$\cos(\angle(\vec{\beta}, \vec{x}_0)) = 1$$

und $\|\beta\| = 1$ (weil wir ja die Hesse Normalform annehmen) und somit

$$\begin{aligned}\langle \vec{\beta}, \vec{x}_0 \rangle - \beta_0 &= 0 \\ \Leftrightarrow \|\vec{\beta}\| \cdot \|\vec{x}_0\| \cdot \cos(\angle(\vec{\beta}, \vec{x}_0)) &= \beta_0 \\ \Leftrightarrow \|\vec{x}_0\| &= \beta_0\end{aligned}$$



Daraus folgt, dass β_0 der Abstand der Ebene zum Ursprung ist.

Hesse Normalform

Für die Hesse Normalform muss $\|\vec{\beta}\| = 1$ gelten. Dann kann aber der Abstand zum Ursprung leicht abgelesen werden. Wir normieren den Normalenvektor der Ebenengleichung also einfach auf die Euklidische Länge 1

$$\vec{\beta} = \frac{\vec{\beta'}}{\|\vec{\beta'}\|}$$

und erhalten die Ebenengleichung in **Hesse Normalform**

$$\langle \vec{\beta}, \vec{x} \rangle - \beta_0 = 0 \quad (2)$$

wobei

$$\beta_0 = \langle \vec{\beta}, \vec{a} \rangle > 0$$

Dann ist β_0 der Abstand zum Ursprung.

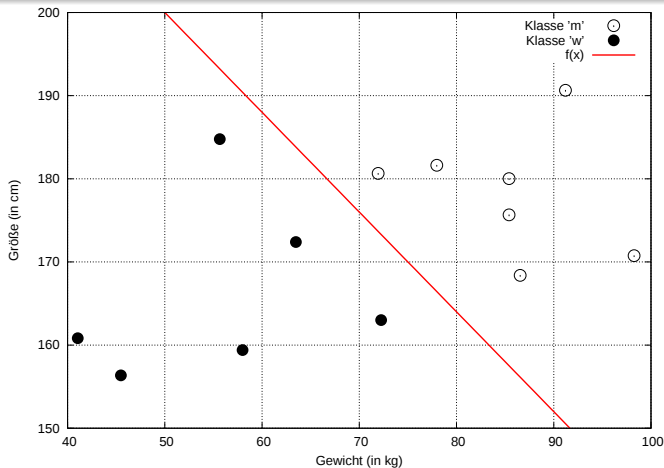
Beispiel Normalisierung

Sei $\vec{a} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$ und $\vec{\beta}' = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ dann ist die Ebenengleichung nicht in Hesse Normalform, weil $\|\vec{\beta}'\| = \sqrt{14} \neq 1$. Wir normalisieren

$$\vec{\beta} = \frac{\vec{\beta}'}{\|\vec{\beta}'\|} = \frac{1}{\sqrt{14}} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
$$\langle \vec{\beta}, \vec{x} \rangle - \beta_0 = 0 \quad \frac{1}{\sqrt{14}}x_1 + \frac{1}{\sqrt{14}}x_2 + \frac{1}{\sqrt{14}}x_3 - \frac{4}{\sqrt{14}} = 0$$

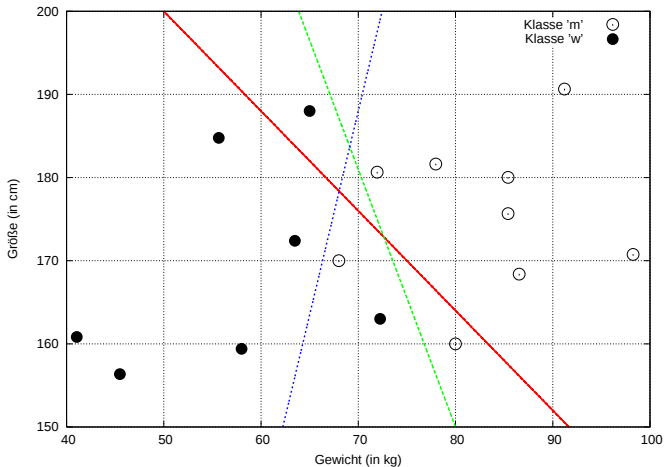
Jetzt ist $\beta_0 = \frac{-4}{\sqrt{14}}$ der Abstand der Ebene zum Ursprung.

Beispiel: Ein mögliches $\vec{\beta}$



$$f(\vec{x}) = \vec{x}^T \hat{\vec{\beta}} \quad \text{mit} \quad \hat{\vec{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 260 \\ 1 \\ 1.2 \end{pmatrix} \quad \theta = 550$$

Es ist nicht garantiert, dass $\vec{\beta}$ immer passt! Es gibt verschiedene Hyperebenen,
;-)



Modell-Anpassung

Unsere linearen Modelle sind durch $\vec{\beta}$ parametrisiert, das Lernen eines Modells haben wir also auf die Wahl eines $\vec{\beta}$ abgewälzt.

Das wirft eine Reihe von Fragen auf:

- Was ist ein gutes $\vec{\beta}$?
- Gibt es ein optimales $\vec{\beta}$?
- Welche Möglichkeiten haben wir, unser Modell zu beurteilen?

Eine Möglichkeit: Berechne den **Trainingsfehler**

$$Err(\vec{\beta}) = \sum_{i=1}^N |y_i - \hat{f}(\vec{x}_i)| = \sum_{i=1}^N |y_i - \mathbf{x}_i^T \vec{\beta}|$$

Modell-Anpassung

Häufig wird hier aber eine andere Fehlerfunktion, nämlich die *quadratische Fehlersumme* (RSS) verwendet:

$$\begin{aligned}RSS(\vec{\beta}) &= \sum_{i=1}^N (y_i - \vec{x}_i^T \vec{\beta})^2 \\ &= (\vec{y} - \mathbf{X}\vec{\beta})^T (\vec{y} - \mathbf{X}\vec{\beta})\end{aligned}$$

Wir wählen jetzt $\vec{\beta}$ derart, dass der Fehler minimiert wird:

$$\min_{\vec{\beta} \in \mathbb{R}^p} RSS(\vec{\beta}) \tag{3}$$

⇒ Konvexes Minimierungsproblem!

Aber das kennt ihr ja schon!!!!

Minimierung von $RSS(\vec{\beta})$

Um $RSS(\vec{\beta})$ zu minimieren, bilden wir die partielle Ableitung nach $\vec{\beta}$:

$$\frac{\partial RSS(\vec{\beta})}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\vec{\beta})$$

Notwendige Bedingung für die Existenz eines (lokalen) Minimums von RSS ist

$$\frac{\partial RSS(\vec{\beta})}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\vec{\beta}) = 0$$

Ist $\mathbf{X}^T\mathbf{X}$ regulär, so erhalten wir

$$\hat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad (4)$$

Reguläre Matrix

Wenn es zu einer quadratischen Matrix $\mathbf{X}$ eine Matrix $\mathbf{X}^{-1}$ gibt mit

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}$$

Einheitsmatrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

dann ist die Matrix $\mathbf{X}$ invertierbar oder *regulär*, sonst *singulär*.

Optimales $\hat{\beta}$?

Mit Hilfe der Minimierung der (quadratischen) Fehlerfunktion RSS auf unseren Trainingsdaten haben wir ein (bzgl. RSS) optimales $\hat{\beta}$ gefunden.

Bei einem konvexen Problem ist das lokale auch das globale Minimum.

Damit liefert unser Modell Voraussagen $\hat{y}$ für $\vec{x} \in X$:

$$\hat{y} = \hat{f}(\vec{x}) = \vec{x}^T \hat{\beta}$$

Sind wir also schon fertig? Mission completed?

- Schön wär's!
- Aber drei Gründe sprechen dagegen:
 - 1 Es ist nicht immer so einfach wie in unserem Beispiel ($p < n$), z.B. dann nicht, wenn wir viele Dimensionen haben (Fluch der hohen Dimension), insbesondere für $p > n$. Hier kann es viele, viele Hyperebenen geben, die die Trainingsdaten separieren.
 - 2 Vielleicht lassen sich die Beispiele nicht linear trennen!
 - 3 Nur den Fehler zu minimieren reicht nicht aus, wir suchen noch nach weiteren Beschränkungen, die zu besseren Lösungen führen.
- Also schauen wir uns den Fehler noch einmal genauer an, stoßen auf Bias und Varianz und merken, dass wir noch keine perfekte Lösung haben.

Fehler

- Bisher haben wir mit RSS die Fehler einfach summiert.
- Wir wollen aber einbeziehen, wie wahrscheinlich der Fehler ist – vielleicht ist er ja ganz unwahrscheinlich! Das machen wir über den Erwartungswert.
- Wir können sehr unterschiedliche Stichproben als Beispielmenge haben. Der Fehler soll sich auf alle möglichen Trainingsmengen beziehen – nicht nur eine, zufällig günstige!

Zur Erinnerung: Erwartungswert

Erwartungswert

Sei X eine **diskrete Zufallsvariable**, mit Werten $x_1, \dots, x_n$ und p_i die Wahrscheinlichkeit für x_i . Der Erwartungswert von X ist

$$E(X) = \sum_i x_i p_i = \sum_i x_i P(X = x_i)$$

Ist X eine **stetige Zufallsvariable** und f die zugehörige Wahrscheinlichkeitsdichtefunktion, so ist der Erwartungswert von X

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Erwartungswert (Eigenschaften)

Eigenschaften

Seien X , Y und $X_1, \dots, X_n$ Zufallsvariablen, dann gilt:

- Der Erwartungswert ist additiv, d.h. es gilt

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) \quad (5)$$

- Ist $Y = kX + d$, so gilt für den Erwartungswert

$$E(Y) = E(kX + d) = kE(X) + d \quad (6)$$

- Sind die Zufallsvariablen X_i **stochastisch unabhängig**, gilt

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$$

Varianz und Standardabweichung

Über den Erwartungswert einer Zufallsvariablen X sind mehrere Eigenschaften von X definiert, die helfen, X zu charakterisieren:

Varianz

Sei X eine Zufallsvariable mit $\mu = E(X)$. Die **Varianz** $Var(X)$ ist definiert als

$$Var(X) := E((X - \mu)^2).$$

Die Varianz wird häufig auch mit σ^2 bezeichnet.

Standardabweichung

Die **Standardabweichung** σ einer Zufallsvariable X ist definiert als

$$\sigma := \sqrt{Var(X)}$$

Varianz und Standardabweichung

Verschiebungssatz

Sei X eine Zufallsvariable, für die Varianz gilt

$$\text{Var}(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2$$

Vereinfacht oft Berechnungen!!!!

Bias

Eine weitere Charakteristik, die häufig zur Beschreibung von erwarteten Fehlern verwendet wird, ist die **Verzerrung**:

Verzerrung (Bias)

Sei Y eine Zufallsvariable, dann ist die Verzerrung definiert als der erwartete Schätzfehler für Y , also **wie im Durchschnitt die Schätzungen vom wahren Mittelwert abweichen**

$$\text{Bias}(\hat{y}) = E(Y - \hat{y}) = E(Y) - \hat{y}$$

“Gute” Schätzer sollten zumindest näherungsweise erwartungstreu, sich also dadurch auszeichnen sollen, dass sie im Mittel nah am zu schätzenden Wert liegen. Sie haben eine geringe Verzerrung

Fehler der Regression

- Fehlerfunktion $L(y, \hat{y})$ für gelernte Modelle $\hat{f}$
 - absolut $\sum (y_i - \hat{y}_i)$
 - quadratisch $\sum (y_i - \hat{y}_i)^2$
 - 0,1-Fehler $\sum \delta_i$, $\delta = 1$, falls $y \neq \hat{y}$, sonst 0.
- Es geht um Y . Wir unterscheiden
 - das wahre y ,
 - das in der Beispielmenge genannte y ,
 - das vom Modell vorhergesagte $\hat{y}$
- Wir wollen den Erwartungswert des Fehlers minimieren.
- Wir mitteln über alle möglichen Beispielmengen $\mathcal{T}$.

Erwartungswert des Fehlers einer Regression minimieren!

Erwarteter quadratischer Vorhersagefehler: Gelernte Funktion $\hat{f} : X \rightarrow Y$, der Erwartungswert ihres Fehlers ist:

$$EPE(f) = E(Y - \hat{f}(X))^2 \quad (7)$$

Optimierungsproblem: Wähle $\hat{f}$ so, dass der erwartete Fehler minimiert wird!

$$\hat{f}(x) = \operatorname{argmin}_c E_{Y|X}((Y - c)^2 | X = x) \quad (8)$$

Lösung (Regressionsfunktion): $\hat{f}(x) = E(Y|X = x)$

Bias und Varianz

Zwei Aspekte machen den erwarteten Fehler aus, die **Verzerrung (Bias)** und die **Varianz**. Wir wollen den Fehler an einem Testpunkt $x_0 = 0$ angeben und mitteln über allen Trainingsmengen $\mathcal{T}$.

- Wir gehen davon aus, dass die Angaben in den Daten nicht immer ganz stimmen, so dass es einen Messfehler ϵ gibt, dessen Erwartungswert aber 0 ist.
- Der Bias ist unabhängig vom Beispielsatz und 0 bei einem perfekten Lerner.
- Die Varianz ist unabhängig vom wahren Wert y und 0 bei einem Lerner, der bei allen Beispielsätzen dasselbe ausgibt.

MSE Dekomposition in Bias und Varianz

Wir nehmen für unser Modell an, dass $Y = f(x) + \epsilon$ und $E(\epsilon) = 0$.
Dann können wir den erwarteten quadratischen Vorhersagefehler wie folgt “erklären”:

$$\begin{aligned} EPE(x_0) &= E_{Y,\mathcal{T}}((Y - \hat{y}_0)^2 | x_0) \\ &= E_Y((Y - f(x_0))^2 | x_0) + \sigma^2 \text{Rauschen} \\ &\quad E_{\mathcal{T}}((f(x_0) - E_{\mathcal{T}}(\hat{y}_0))^2 | x_0) + \text{Bias}^2 \\ &\quad E_{\mathcal{T}}((E_{\mathcal{T}}(\hat{y}_0) - \hat{y}_0)^2 | x_0) \quad \text{Varianz} \end{aligned}$$

Aber wie kommen wir darauf?!

Haupttrick: kreatives Einfügen von Termen, $+a - a$, die nichts ändern, aber Umformungen erlauben.

Herleitung der Dekomposition - 1

$$\begin{aligned} EPE(x_0) &= E_{Y,\mathcal{T}}((Y - \hat{y}_0)^2|x_0) \\ &= E_{Y,\mathcal{T}}(((Y - f(x_0)) + (f(x_0) - \hat{y}_0))^2|x_0) && \textit{kreativ} \\ &= E_{Y,\mathcal{T}}((Y - f(x_0))^2 + (f(x_0) - \hat{y}_0)^2 + && \textit{binomisch} \\ &\quad 2(Y - f(x_0))(f(x_0) - \hat{y}_0)|x_0) \\ &= E_Y((Y - f(x_0))^2|x_0) + && \textit{herausziehen} \\ &\quad E_{Y,\mathcal{T}}((f(x_0) - \hat{y}_0)^2|x_0) + && \textit{s.Formel(5)} \\ &\quad 2E_Y(Y - f(x_0)|x_0)E_{Y,\mathcal{T}}(f(x_0) - \hat{y}_0|x_0) && E(2) = 2 \end{aligned}$$

Jetzt $E_Y(Y - f(x_0)|x_0) = E_Y(Y|x_0) - f(x_0) = 0$ wegen der Annahme $E_Y(Y|x_0) = f(x_0)$.

Herleitung der Dekomposition - 2

$$\begin{aligned}
 EPE(x_0) &= E_Y((Y - f(x_0))^2|x_0) + && Var_{\epsilon} \\
 &E_{Y,T}((f(x_0) - \hat{y}_0)^2|x_0) \\
 &= Var_{\epsilon} + E_{Y,T}((f(x_0) - \hat{y}_0)^2|x_0) \\
 &= Var_{\epsilon} + E_{Y,T}(((f(x_0) - E_{Y,T}(\hat{y}_0)) + && kreativ \\
 &(E_{Y,T}(\hat{y}_0) - \hat{y}_0))^2|x_0) \\
 &= Var_{\epsilon} + E_{Y,T}((f(x_0) - E_{Y,T}(\hat{y}_0))^2 + && binomisch \\
 &(E_{Y,T}(\hat{y}_0) - \hat{y}_0)^2 + \\
 &2(f(x_0) - E_{Y,T}(\hat{y}_0))(E_{Y,T}(\hat{y}_0) - \hat{y}_0|x_0)) \\
 &= Var_{\epsilon} + E_{Y,T}((f(x_0) - E_{Y,T}(\hat{y}_0))^2|x_0) + && herausziehen \\
 &E_{Y,T}((E_{Y,T}(\hat{y}_0) - \hat{y}_0)^2|x_0) + && (5) \\
 &2E_{Y,T}((f(x_0) - E_{Y,T}(\hat{y}_0)) \\
 &(E_{Y,T}(\hat{y}_0) - \hat{y}_0)|x_0)
 \end{aligned}$$

Herleitung der Dekomposition - 3

Rauschen haben wir schon gefunden. Bias und Varianz aber auch schon !

$$\begin{aligned}
 EPE(x_0) = & \quad Var_{\epsilon} + \quad \sigma^2 \text{Rauschen} \\
 & E_{Y,\mathcal{T}}((f(x_0) - E_{Y,\mathcal{T}}(\hat{y}_0))^2 | x_0) + \quad Bias^2 \\
 & E_{Y,\mathcal{T}}((E_{Y,\mathcal{T}}(\hat{y}_0) - \hat{y}_0)^2 | x_0) + \quad Varianz \\
 & 2E_{Y,\mathcal{T}}(f(x_0) - E_{Y,\mathcal{T}}(\hat{y}_0)) \quad Konstanten \\
 & (E_{Y,\mathcal{T}}(E_{Y,\mathcal{T}}\hat{y}_0 - \hat{y}_0 | x_0)) \\
 = & \mathcal{O}(\text{Rauschen} + Bias^2 + Varianz)
 \end{aligned}$$

Wir betrachten den Fehler also als zusammengesetzt aus Rauschen, Verzerrung und Varianz. Die Dekomposition des MSE in Bias und Varianz abstrahiert so, dass wir besser über Modelle nachdenken können.

Bias und Varianz bei linearen Modellen

Das lineare Modell wird an die Daten angepasst durch

$$\hat{f}_p(\vec{x}) = \hat{\beta}^T \vec{x}$$

Der Fehler ist dann für ein beliebiges $\vec{x}$:

$$Err(\vec{x}_0) = E[(Y - \hat{f}_p(\vec{x}_0))^2 | X = \vec{x}_0] \quad (9)$$

$$= \sigma_\epsilon^2 + Var(\hat{f}_p(\vec{x}_0)) + \left[f(\vec{x}_0) - E\hat{f}_p(\vec{x}_0) \right]^2 \quad (10)$$

Die Anpassung des linearen Modells geht über alle N Beispiele und gewichtet alle p Merkmale (s. (4)).

Diese Varianz ist von x_i zu x_i verschieden. Im Mittel über allen $\vec{x}_i$ ist $Var(\hat{f}_p) = (p/N)\sigma_\epsilon^2$.

Zusammenhang zwischen Anzahl der Beispiele, der Attribute und erwartetem Fehler

Modellkomplexität (p, N) und **Varianz** der Schätzungen bei unterschiedlichen Trainingsmengen hängen bei linearen Modellen direkt zusammen.

Gemittelt über alle x_i ist der Trainingsfehler linearer Modelle:

$$\frac{1}{N} \sum_{i=1}^N \text{Err}(x_i) = \sigma_{\epsilon}^2 + \frac{p}{N} \sigma_{\epsilon}^2 + \frac{1}{N} \sum_{i=1}^N \left[f(\vec{x}_i) - E\hat{f}(\vec{x}_i) \right]^2 \quad (11)$$

Wir haben also wieder das Rauschen, die Varianz, die die Schwankungen der Schätzungen angibt, und den Bias, der sich auf die Differenz von Schätzung und Wahrheit bezieht (in-sample error).

Fluch der hohen Dimension bei linearen Modellen

- Leider mussten wir annehmen, dass das Modell genau passt, um den erwarteten Fehler klein zu halten.
- Wir wissen aber nicht, welche Art von Funktion gut zu unseren Daten passt! **Modellselektion** ist schwierig!
- Das Modell muss immer komplizierter werden, je mehr Dimensionen es gibt.
- Bei linearen Modellen entspricht die Komplexität des Modells direkt p , denn β hat so viele Komponenten wie p bzw. $p + 1$.

Zusammenhang Bias, Varianz und Modellkomplexität – bildlich

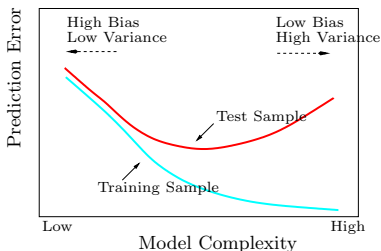


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

Wir haben also einen Tradeoff zwischen Bias und Varianz:

- Je komplexer wir das Modell machen, desto geringer der quadrierte Bias und desto größer gleichzeitig die Varianz.
- Umgekehrt weisen einfachere Modelle einen größeren quadrierten Bias und im Ausgleich eine kleinere Varianz auf.

Was wissen Sie jetzt?

- Sie haben theoretisch lineare Modelle für Klassifikation und Regression kennengelernt.
- Sie kennen den erwarteten Fehler EPE bei linearen Modellen (Gleichung 7).
- Sie kennen den **Fluch der hohen Dimension** bei linearen Modellen: Komplexität und Varianz hängen an der Dimension! Der Bias kann sehr hoch sein, wenn die Beispiele tatsächlich nicht linear separierbar sind.

Bis zum nächsten Mal...

- Gehen Sie alle Folien noch einmal in Ruhe durch.
- Vertiefen Sie sich noch einmal in die Ebenengleichung (2)! Die lineare Algebra wird immer wieder vorkommen. Sie können auch die partiellen Ableitungen für RSS mit der Normalengleichung vornehmen.
- Rechnen Sie mal ein Beispiel durch mit Gleichung zur Optimierung linearer Modelle (4), der Minimierung des Trainingsfehlers (11)...
- Diskutieren Sie, warum Bias und Varianz so wichtig sind!
- Vielleicht probieren Sie lineare Regression in RapidMiner aus!