



Vorlesung Maschinelles Lernen

Klassifikation und Regression: Lineare Modelle

Katharina Morik

LS 8 Künstliche Intelligenz Fakultät für Informatik
Technische Universität Dortmund

25.10.2008



Gliederung

- 1 **Lineare Modelle zur Klassifikation und Regression**
 - Klassifikation und Regression
 - Lineare Modelle
 - Geometrie linearer Modelle: Hyperebenen

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



Grundlagen

Sei $X = \{X_1, \dots, X_p\}$ eine Menge von 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 Matrix \mathbf{X} einem Beispiel.



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 $(\vec{x}, y) \in X \times Y$.

Y kann sowohl eine **qualitative**, als auch eine **quantitative** Beschreibung von \vec{x} sein.

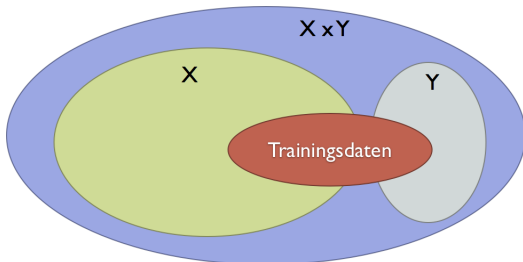
Für den quantitativen Fall ist z.B. $Y = \mathbb{R}$ und wir versuchen für unbekanntes \vec{x} den Wert y vorherzusagen **Regression**.

Im Falle qualitativer Beschreibungen ist Y eine diskrete Menge und wir nutzen f zur **Klassifikation**.



Lernen auf Trainingsdaten

Wovon gehen wir also 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.

Dieses Modell \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

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

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

Die Tabelle enthält die zur Verfügung stehenden Trainingsdaten, also

$$\mathbf{X} = \begin{pmatrix} 91 & 190 \\ 60 & 170 \\ 41 & 160 \\ \vdots & \vdots \end{pmatrix}$$



Beispiel

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

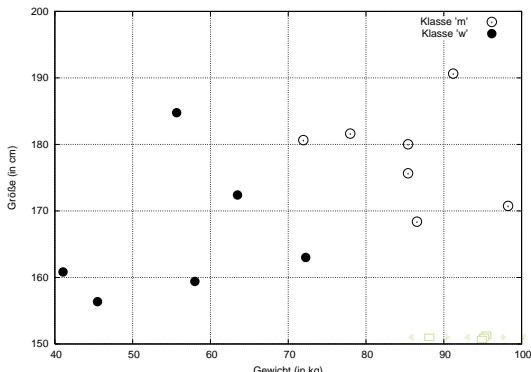
$$\hat{y} = \begin{cases} m, & \text{falls } \hat{f}(x) > \theta \\ w, & \text{sonst} \end{cases}$$



Beispiel

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

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





Lineare Modelle

Welche Art von Funktionen sind denkbar?

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:

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



Was haben wir nun gemacht?

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

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

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



Veranschaulichung

Bevor wir uns an die Wahl des passenden $\vec{\beta}$ machen, zunächst einige Vorüberlegungen.

Betrachten wir dazu die binäre Klassifikation ($Y = \{-1, +1\}$):

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



Zur Erinnerung: Ebenengleichung

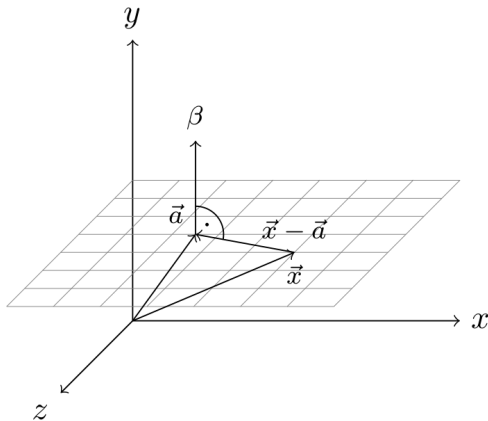
Sei $V = \mathbb{R}^p$ ein Vektorraum, dann ist eine Hyperebene H ein $(p - 1)$ -dimensionaler affiner Untervektorraum.

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,
 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

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

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

aber auch durch den Kosinus mit

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



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

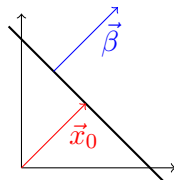
$$\cos(\angle(\vec{\beta}, \vec{x}_0)) = 1$$

und $\|\beta\| = 1$ und somit

$$\langle \vec{\beta}, \vec{x}_0 \rangle - \beta_0 = 0$$

$$\Leftrightarrow \|\vec{\beta}\| \cdot \|\vec{x}_0\| \cdot \cos(\angle(\vec{\beta}, \vec{x}_0)) = \beta_0$$

$$\Leftrightarrow \|\vec{x}_0\| = \beta_0$$



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



Hesse Normalform

Für die Hesse Normalform muss $\|\vec{\beta}\| = 1$ gelten, damit der Abstand zum Ursprung leicht abgelesen werden kann. Wir normieren den Normalenvektor 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

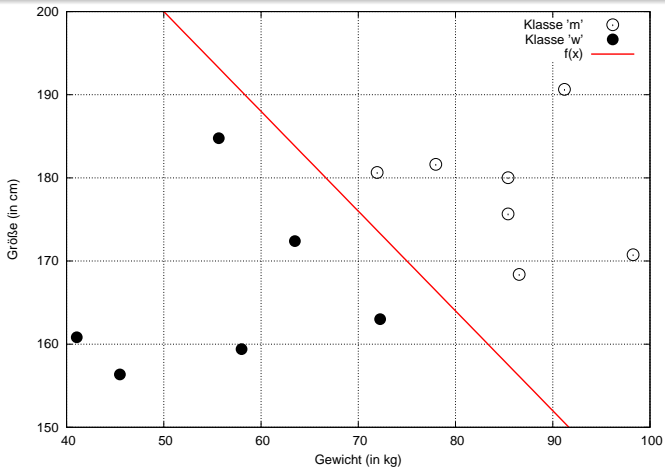
$$\vec{\beta} = \frac{\vec{\beta}'}{\|\vec{\beta}'\|} = \frac{1}{\sqrt{14}} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

$$\langle \vec{\beta}, \vec{x} \rangle - \beta_0 = 0 \quad \frac{1}{\sqrt{14}}x_1 + \frac{1}{\sqrt{14}}x_2 + \frac{1}{\sqrt{14}}x_3 - \frac{4}{\sqrt{14}} = 0$$

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



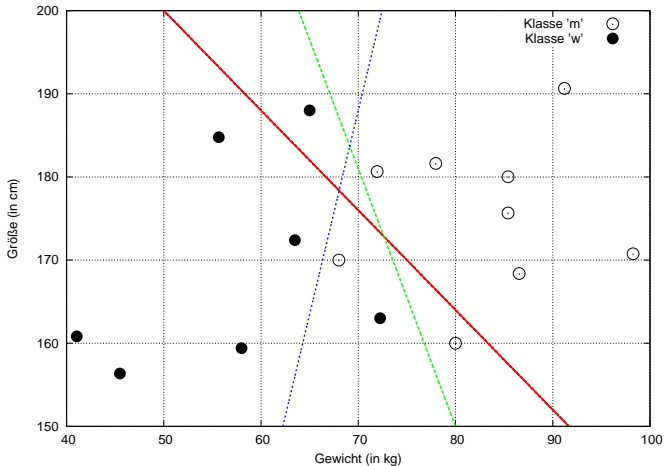
Beispiel: Ein mögliches $\vec{\beta}$



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



Es ist nicht garantiert, dass $\vec{\beta}$ immer passt!





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?



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*

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



Modell-Anpassung

Häufig wird als Fehlerfunktion 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:

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

⇒ Konvexes Minimierungsproblem!

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

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & \cdot & \dots & 0 \\ \cdot & \cdot & \dots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

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



Optimales $\hat{\vec{\beta}}$?

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

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

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

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



Sind wir schon fertig?

- Schön wär's!
- Aber drei Gründe sprechen für weitere Arbeit:
 - 1 Es ist nicht immer so einfach, z.B. dann nicht, wenn wir viele Dimensionen haben (Fluch der hohen Dimension).
 - 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!
- Wann können wir denn einen Fehler erwarten?



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

$$E(Y) = E(kX + d) = kE(X) + d$$

- 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

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

Standardabweichung

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

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



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

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

Standardabweichung

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

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

Die Varianz wird häufig auch mit σ^2 bezeichnet.



Varianz und Standardabweichung

Verschiebungssatz

Sei X eine Zufallsvariable, für die Varianz gilt

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

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

Verzerrung (Bias)

Sei X eine Zufallsvariable, dann ist die Verzerrung definiert als der erwartete Schätzfehler für X

$$\text{Bias}(\hat{x}) = E(X - \hat{x})$$



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

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

Satz: Ist die Dichtefunktion f einer Zufallsvariablen symmetrisch um einen Wert c , also

$$f(c - x) = f(c + x) \quad \forall x$$

dann ist $E(X) = c$.



Erwartungswert des Fehlers einer Regression

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

$$EPE(f) = E(Y - \hat{f}(X))^2 \quad (6)$$

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

$$\hat{f}(x) = \operatorname{argmin}_c E_{Y|X}((Y - c)^2 | X = x) \quad (7)$$

Lösung (Regressionsfunktion): $f(x) = E(Y|X = x)$



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} .

$$\begin{aligned}MSE(\vec{x}_0) &= E_{\mathcal{T}}[y_0 - \hat{y}_0]^2 \\&= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)]^2 + [E_{\mathcal{T}}(\hat{y}_0 - y_0)]^2 \\&= E_{\mathcal{T}}[y_0^2] - E_{\mathcal{T}}[2y_0\hat{y}_0] + E_{\mathcal{T}}[\hat{y}_0^2] \\&= \text{Var}_{\mathcal{T}}(\hat{y}_0) + \text{Bias}^2(\hat{y}_0)\end{aligned}$$



Erwartungswert des Fehlers bei linearen Modellen

Unter der **Annahme**, dass unsere Beispiele Messfehler enthalten, aber X und Y wirklich linear voneinander abhängen (**Bias=0**), passen wir das Modell $Y = X^T \beta + \epsilon$ durch Minimieren des quadratischen Fehlers an.

Der erwartete Fehler der \hat{y} -Vorhersage für ein beliebiges \vec{x}_0 ist:

$$\begin{aligned} EPE(\vec{x}_0) &= E_{y_0|\vec{x}_0} E_{\mathcal{T}}(y_0 - \hat{y}_0)^2 \\ &= \text{Var}(y_0|\vec{x}_0) + E_{\mathcal{T}}(\hat{y}_0 - E_{\mathcal{T}}(y_0))^2 + (E_{\mathcal{T}}(\hat{y}_0) - E_{\mathcal{T}}(y_0))^2 \\ &= \text{Var}(y_0|\vec{x}_0) + \text{Var}_{\mathcal{T}}(\hat{y}_0) + \text{Bias}^2(\hat{y}_0) \\ &= \sigma^2 + E_{\mathcal{T}}(\vec{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \vec{x}_0 \sigma^2) + 0^2 \end{aligned}$$

Die zusätzliche Varianz kommt durch das Rauschen.



Zusammenhang zwischen Anzahl der Beispiele, der Attribute und erwartetem Fehler

Beziehen wir den Erwartungswert von \vec{x} ein, erhalten wir

$$E_{\vec{x}} EPE(\vec{x}) = \sigma^2(p/N) + \sigma^2$$

Bei kleinem σ^2 und großem N klappt alles auch bei großem p ,
wenn das lineare Modell perfekt passt, d.h. die Beispiele sind
linear trennbar.



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. $p + 1$.



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} :

$$Err(\vec{x}) = E[(Y - \hat{f}_p(\vec{x}))^2 | X = \vec{x}] \quad (8)$$

$$= \sigma_\epsilon^2 + Var(\hat{f}_p(\vec{x})) + \left[f(\vec{x}) - E\hat{f}_p(\vec{x}) \right]^2 \quad (9)$$

Im Mittel über allen \vec{x}_i ist $Var(\hat{f}_p) = (p/N)\sigma^2$. Modellkomplexität und Varianz hängen bei linearen Modellen direkt zusammen.

Der Trainingsfehler linearer Modelle ist:

$$\frac{1}{N} \sum_{i=1}^N Err(x_i) = \sigma_\epsilon^2 + \frac{p}{N} \sigma_\epsilon^2 \frac{1}{N} \sum_{i=1}^N \left[f(\vec{x}_i) - E\hat{f}(\vec{x}_i) \right]^2 \quad (10)$$



Lineare Modelle

Die grünen und roten Datenpunkte werden durch eine Ebene getrennt.

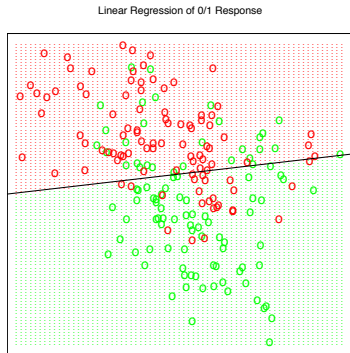


Figure 2.1: A classification example in two dimensions. The classes are coded as a binary variable—**GREEN** = 0, **RED** = 1—and then fit by linear regression. The line is the decision boundary defined by $x^T \hat{\beta} = 0.5$.



Was wissen Sie jetzt?

- Sie haben theoretisch lineare Modelle für Klassifikation und Regression kennengelernt.
- Sie kennen das **Optimierungsproblem** der kleinsten Quadrate RSS (Gleichung 3) für lineare Modelle (Gleichung 4).
- Sie kennen den erwarteten Fehler EPE bei linearen Modellen (Gleichung 6).
- 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 (10)...
- Diskutieren Sie, warum Bias und Varianz so wichtig sind!
- Probieren Sie lineare Regression in RapidMiner aus!