Künstliche Intelligenz / Maschinelles Lernen

Klassifikation, Regression und Lineare Modelle

TECHNISCHE UNIVERSITÄT DARMSTADT

Lineare Modelle sind statistisches Modelle, bei denen der Erwartungswert einer Variable Y in einer bestimmten ("linearen") Weise von Eingabevariablen $\mathbf{X}$ abhängt. Sie versuchen also den Zusammenhang zwischen einer abhängigen Variablen (oder Responsevariablen) Y und einer oder mehreren erklärenden Variablen $X_1, \ldots, X_k$ zu modellieren. In R kann z.B. `lm()` benutzt werden.

Basierend auf Folien von Katharina Morik und Uwe Ligges. Danke fürs Offenlegen der Folien.

Unsere Ziele

TECHNISCHE UNIVERSITÄT DARMSTADT

Was wollen wir hier kennenlernen?

- Was versteht man unter Klassifikation?
- Was versteht man unter Regression?
- Was sind lineare Modelle?
- Wie bestimmt ich lineare Modelle aus Daten?
- Wie evaluiere ich Modelle auf Daten?

Grundlagen

TECHNISCHE UNIVERSITÄT DARMSTADT

Sei $X = \{X_1, \ldots, X_p\}$ 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 $\mathbf{X} = \{\vec{x}_1, \ldots, \vec{x}_N\}$ können wir dann als $(N \times p)$-Matrix auffassen:

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

Dabei entspricht jede Zeile $\vec{x}_i$ der Matrix $\mathbf{X}$ einem Beispiel.

Klassifikation und Regression

TECHNISCHE UNIVERSITÄT DARMSTADT

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.

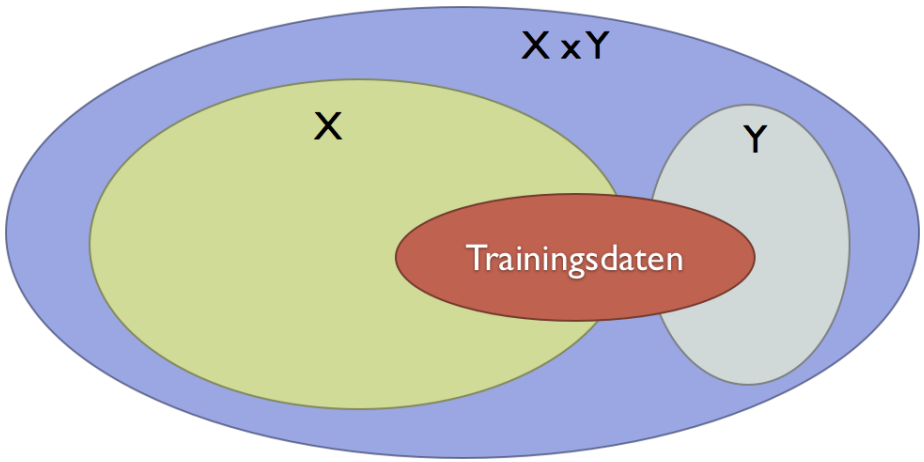
Für den quantitativen Fall ist z.B. $Y = \mathbb{R}$ und wir versuchen für unbekanntes $\vec{x}$ den Wert y vorherzusagen Regression.

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

Lernen auf Trainingsdaten

TECHNISCHE UNIVERSITÄT DARMSTADT

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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

Klassifikation und Regression

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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:

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

Hier ist θ ein vorgegebener Schwellwert.

Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT

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

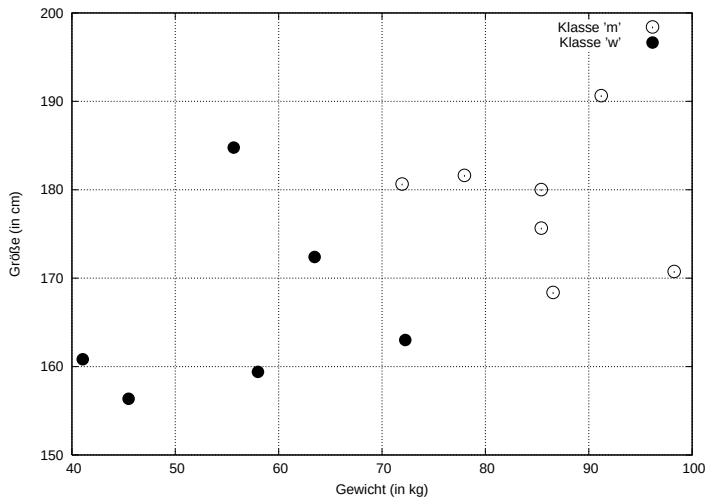
	X_1	X_2	Y
x_1	91	190	m
x_2	60	170	w
x_3	41	160	w
⋮	⋮	⋮	⋮

Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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



Lineare Modelle

TECHNISCHE UNIVERSITÄT DARMSTADT

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:

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

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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

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

Dies ermöglicht die Darstellung von f als:

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

Was haben wir nun gemacht?

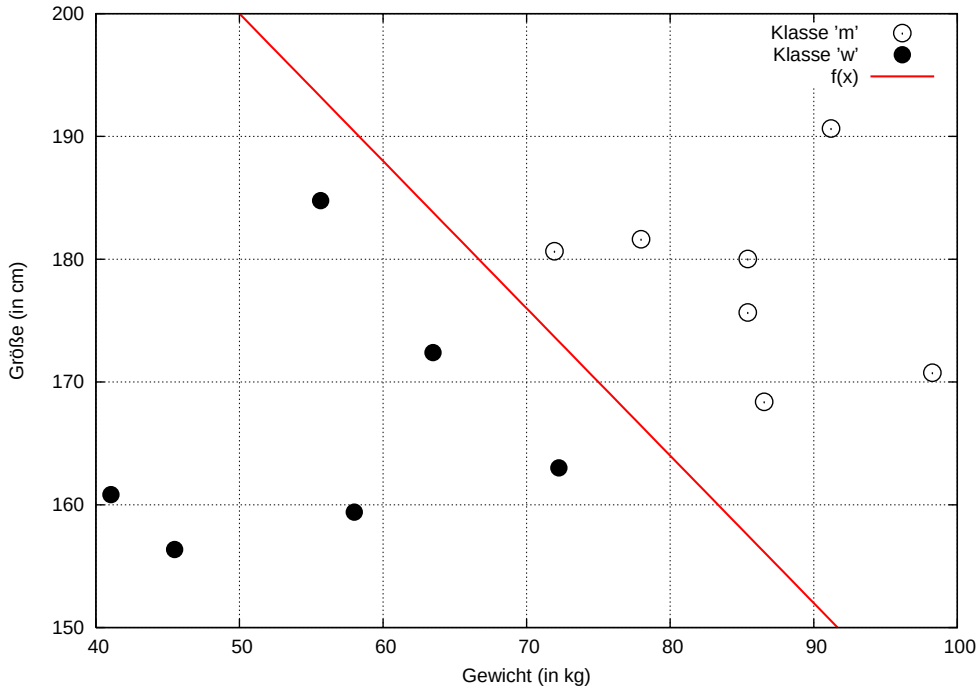
TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

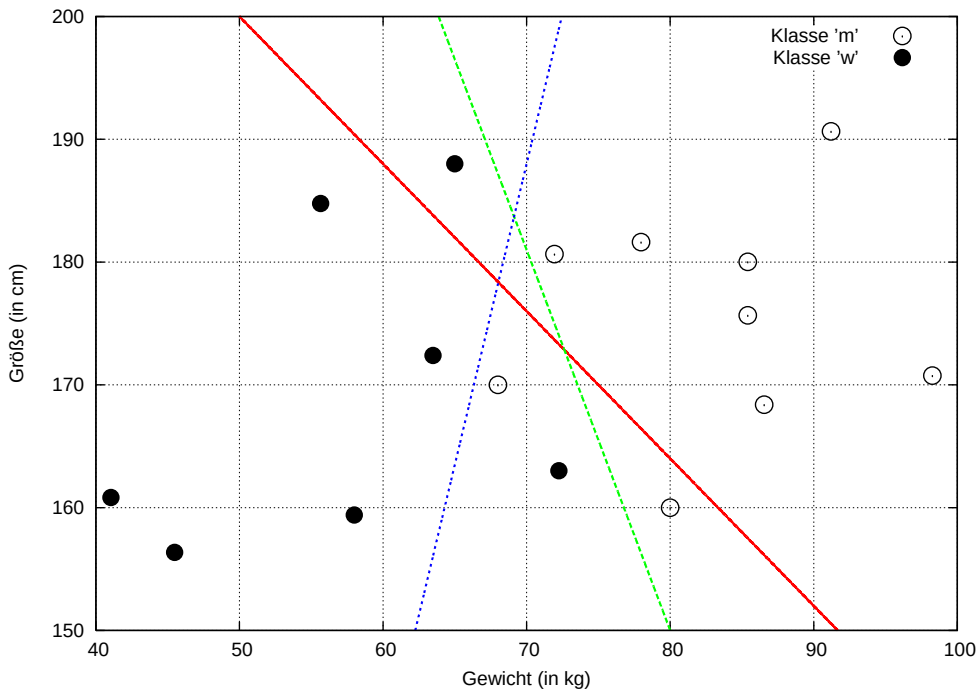
Wir haben die Zielfunktion $\hat{f}$ in Abhängigkeit von $\vec{\beta}$ geschrieben und müssen *nur noch* das passende $\vec{\beta}$ finden.

Beispiel: Ein mögliches $\vec{\beta}$



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

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



Modell-Anpassung

TECHNISCHE UNIVERSITÄT DARMSTADT

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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:

$$\min_{\vec{\beta} \in \mathbb{R}^p} RSS(\vec{\beta}) \tag{2}$$

$\Rightarrow$ Konvexes (="einfaches") Minimierungsproblem!

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

$$\hat{\vec{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

Reguläre Matrix

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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 $\vec{x} \in X$:

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

Sind wir schon fertig?

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

Erwartungswert

Sei X eine diskrete Zufallsvariable, mit Werten $x_1, \ldots, x_n$ 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)

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \tag{4}$$

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

$$E(Y) = E(kX + d) = kE(X) + d \tag{5}$$

- 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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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

TECHNISCHE UNIVERSITÄT DARMSTADT

Verschiebungssatz

Sei X eine Zufallsvariable, für die Varianz gilt

$$Var(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2$$

Bias

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

Fehler der Regression

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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 $y = \hat{y}$, 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!

TECHNISCHE UNIVERSITÄT DARMSTADT

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

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

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

$$\hat{f}(x) = argmin_c E_{Y|X}((Y - c)^2 | X = x) \tag{7}$$

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

Bias und Varianz

TECHNISCHE UNIVERSITÄT DARMSTADT

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.

Dekomposition in Bias und Varianz

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir nehmen für unser Modell an, dass $Y = f(x) + \epsilon$ und $E(\epsilon) = 0$.

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

Wie das?!

Haupttrick: kreatives Einfügen von Termen, $+a - a$, die nichts ändern, aber Umformungen erlauben. Wir leiten das hier aber nicht her.

Bias und Varianz bei linearen Modellen

TECHNISCHE UNIVERSITÄT DARMSTADT

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 $\vec{x}$:

$$Err(\vec{x_0}) = E[(Y - \hat{f}_p(\vec{x_0}))^2 | X = \vec{x_0}] \quad (8)$$

$$= \sigma_\epsilon^2 + Var(\hat{f}_p(\vec{x_0})) + \left[f(\vec{x_0}) - E\hat{f}_p(\vec{x_0})\right]^2 \quad (9)$$

Die Anpassung des linearen Modells geht über alle N Beispiele und gewichtet alle p Merkmale (s. (3)).

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

TECHNISCHE UNIVERSITÄT DARMSTADT

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:

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

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

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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. $p + 1$.

Lineare Modelle

TECHNISCHE UNIVERSITÄT DARMSTADT

Die grünen und roten Datenpunkte werden durch eine Ebene getrennt.

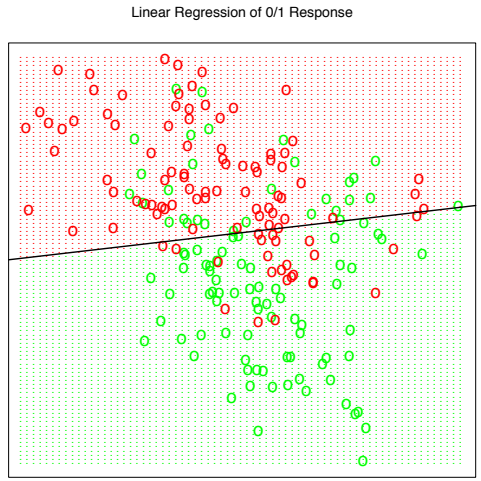


Figure 2.1: *A classification example in two dimensions. The classes are coded as a binary variable—*GREEN $= 0$, RED $= 1$*—and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The red shaded region denotes that part of input space classified as*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

Was wissen Sie jetzt?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sie haben theoretisch lineare Modelle für Klassifikation und Regression kennengelernt.
- Sie kennen das Optimierungsproblem der kleinsten Quadrate RSS (Gleichung 2) für lineare Modelle (Gleichung 3).
- 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.