# Vorlesung 13 Linear Regression

Von letzter VL:

Gegeben sind Daten $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$

Wir haben zwei Arten von Modellen

1. Deterministische Modelle:

$$f_\theta : \mathbb{R}^D \to \mathbb{R}^N, \quad \theta \in \mathbb{R}^P.$$

2. Statistisches Modell:

$$P_\theta(y \mid x)$$

mit $\theta \in \mathbb{R}^P$ und $P_\theta$ die Wkt.-Dichte von $y$ gegeben $x$.

Strategien zur Parameterbestimmung waren.

1. ERM: (für det. Modell).

$$\theta \in \operatorname*{argmin}_{\theta \in \mathbb{R}^P} R(\theta), \quad \text{wobei } R(\theta) = \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$$

Loss-Funktion.

2. MLE: (für stat. Modell)

$$\theta \in \operatorname*{argmax}_{\theta \in \mathbb{R}^P} \prod_{i=1}^{n} P_\theta(y_i \mid x_i)$$

$X = [x_1, \dots, x_n]$
$Y = [y_1, \dots, y_n]$

3. MAP (für stat. Modell)

$$\theta \in \operatorname{argmax} \alpha(\theta), \quad \alpha(\theta) = P(\theta \mid X, Y),$$

wobei $\theta$ selbst als zufällig angesehen.

4. Bayes'sche Ansatz:

Die Wkt-Dichte $P(\theta \mid X Y)$ berechnen und $\theta$

samplen oder $P(y|x,X,Y) = \int_\Theta P_\Theta(y|x) P(\Theta|X,Y) d\Theta$
berechnen

Kurz zurück zur Modellauswahl: Den Bayes'schen Ansatz Parameter selbst als zufällig zu modellieren können wir auch hier benutzen.

Angenommen wir haben Modelle $M_1, \dots, M_K$ zur Auswahl.
Wir setzen den <u>Prior</u> $P(M)$ auf die Menge der Modell (d.h. $P(M)$ bestimmt eine Wkt-Verteilung auf $\{M_1, \dots, M_K\}$). Z.B.: $P(M = M_i) = \frac{1}{K}$.
Dann haben wir die Posterior-Funktion $P(M|X,Y)$
Es gilt nach Bayes' Theorem:
$\operatorname{argmax}_i \; P(M = M_i | X,Y) = \operatorname{argmax}_i \; P(M_i) P(X,Y|M_i)$
und $P(X,Y|M_i) = \int_\Theta P(X,Y|\Theta) \underbrace{P(\Theta|M_i)}_{\text{Prior für Modell } M_i}$

<u>Lineare Regression</u>

<u>Definition</u> 13.1
Das <u>Lineare Modell</u> $f_\Theta : \mathbb{R}^D \to \mathbb{R}$ ist gegeben durch

$$\Theta = (\Theta_0, \Theta_1, \dots, \Theta_D)^T \in \mathbb{R}^{D+1}$$

mit

$$f_\Theta(x) = x^T \Theta' + \Theta_0,$$

wobei $\Theta' = (\Theta_1, \dots, \Theta_D)^T \in \mathbb{R}^D$.
Der <u>quadratische Loss</u> ist

$$\ell(y, \hat{y}) = (y - \hat{y})^2.$$

Das zugehörige ERM heißt das <u>Problem der kleinsten Quadrate</u>

bzw. <u>Least-squares</u>.

Gegeben seien wieder Daten $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$.
Wir setzen wieder

$$X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times D}, \quad Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$$

Wir setzen außerdem

$$\Omega = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}^T \in \mathbb{R}^{n \times (D+1)}$$

$\Omega$ heißt <u>Feature-matrix</u>

<u>Theorem 13.2</u>

Sei $r(\Omega)$ der Rang von $\Omega$. Dann gilt im obigen Setting:

1. Wenn $r(\Omega) < D+1$, hat $\underset{\theta}{\operatorname{argmin}}\, R(\theta)$ unendlich viele Lösungen.
2. Wenn $r(\Omega) = D+1$, hat $\underset{\theta}{\operatorname{argmin}}\, R(\theta)$ eine eindeutige Lösung:

$$\theta_{ERM} = \Omega^+ Y,$$

wobei $\Omega^+$ die Pseudoinverse von $\Omega$ ist.

<u>Beweis</u>

In unserem Fall ist das empirische Risiko:

$$\begin{aligned} R(\theta) &= \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \\ &= \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 \\ &= \sum_{i=1}^{n} (y_i - x_i^T \theta' + \theta_0)^2 \end{aligned}$$

$$= \left\| \begin{pmatrix} y_1 - x_1^T \Theta' + \Theta_0 \\ \vdots \\ y_n - x_n^T \Theta' + \Theta_0 \end{pmatrix} \right\|^2$$

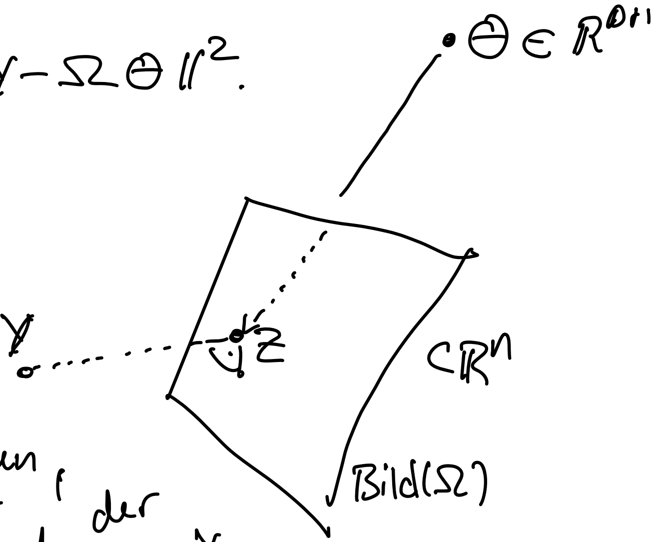
$$= \| Y - \Omega \cdot \Theta \|^2 .$$

D.h. $\Theta_{ERM} = \operatorname{argmin}_{\Theta \in \mathbb{R}^{D+1}} \| Y - \Omega \Theta \|^2$.

Sei $z := \Omega \Omega^+ Y$

Dann wissen wir

$$z = \operatorname*{argmin}_{w \in \text{Bild}(\Omega)} \| Y - w \|^2 .$$




$\Theta$ muss jetzt die Gleichung $z = \Omega \Theta$ erfüllen, weil $\| Y - z \|^2$ der minimale Abstand von $Y$ nach $\text{Bild}(\Omega)$ ist.

Falls $r(\Omega) < D+1$, hat $z = \Omega \Theta$ unendl. viele Lösungen, weil $\ker(\Omega) \neq 0$.

Falls $r(\Omega) = D+1$, haben wir gezeigt, dass $\Omega^+ \Omega = \mathbb{1}_{D+1}$.

$\Rightarrow$ $\Theta = \Omega^+ Y$ eindeutig definiert ist. $\square$

$= \Omega^+ z$

Der Kleinste-Quadrate-Schätzer führt oft zu Overfitting. Um das zu verhindern wird oft ein Regularisierungs-Parameter $\lambda$ eingeführt. Dieser definiert die <u>regularisierte Loss-Funktion</u>

$$\ell(y, \hat{y}) = (y - \hat{y})^2 + \lambda \cdot \| A \Theta \|^2 \quad \text{mit } A \in \mathbb{R}^{(D+1) \times (D+1)} .$$

Der Spezialfall $A = 1_{D+1}$ heißt Tikhonov-Regularisierung und das zugehörige Regressionsproblem heißt <u>Ridge Regression</u>

<u>Theorem</u> 13.3

Sei $\Omega \in \mathbb{R}^{n \times (D+1)}$ die Feature Matrix und $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$.

Für fast alle (= alle außer endlich viele) Werte von $\lambda$ gibt es eine eindeutige Lösung

$$\Theta_{RR} = \underset{\Theta}{\operatorname{argmin}}\, R(\Theta).$$

mit

$$\Theta_{RR} = (\Omega^T\Omega + n\lambda A^T A)^{-1}\Omega^T Y.$$

<u>Bemerkung</u> Thm 13.3. kann auch verwendet werden, wenn $r(\Omega) < D+1$, d.h. z.B. falls $n < D+1$ (weniger Datenpunkte als Parameter).

<u>Beweis von Thm</u> 13.3.

Das empirische Risiko ist:

$$\begin{aligned} R(\Theta) &= \sum_{i=1}^{n} \ell(y_i, f_\Theta(x_i)) \\ &= \sum_{i=1}^{n} (y_i - f_\Theta(x_i))^2 + \lambda \|A\Theta\|^2 \\ &= \|Y - \Omega\Theta\|^2 + n\lambda \|A\Theta\|^2. \end{aligned}$$

$\Theta_{RR}$ minimiert $R(\Theta)$. Um $\Theta_{RR}$ zu finden sehen wir

$$\frac{d}{d\Theta} R(\Theta) = 0.$$

D.h. $$0 = \frac{d}{d\Theta}\left((Y-\Omega\Theta)^T(Y-\Omega\Theta) + n\lambda\, \Theta^T A^T A\Theta.\right)$$

$$= 2\ \Omega^T(Y-\Omega\Theta) + 2n\lambda A^TA\Theta$$

$$\rightarrow \quad \underbrace{(\Omega^T\Omega - n\lambda A^TA)}_{\in \mathbb{R}^{(D+1)\times(D+1)}}\Theta = \Omega^T Y.$$

Es gilt: $\Omega^T\Omega - n\lambda A^TA$ ist invertierbar, falls

$$f(\lambda) = \det(\Omega^T\Omega - n\lambda A^TA) \neq 0.$$

Aber $f(\lambda)$ hat höchstens $D+1$ Nullstellen, weil es ein Polynom vom Grad $D+1$ ist.

$\rightsquigarrow$ Bis auf endlich viele Werte von $\lambda$ ist $\Omega^T\Omega - n\lambda A^TA$ invertierbar und es gilt:

$$\Theta_{RR} = (\Omega^T\Omega - n\lambda A^TA)^{-1}\Omega^T Y.$$

Jetzt: statistisches Modell. Wir wählen folgende Verteilung für $y$ gegeben $x$:

$$P_\Theta(y|x) = \Phi(y \,|\, f_\Theta(x), \sigma^2).$$

D.h. $(y|x) = f_\Theta(x) + \varepsilon$, mit $\varepsilon \sim N(0, \sigma^2)$.

(wieder ist $\sigma^2$ ein Parameter, den wir nicht optimieren, sondern vorher festlegen, d.h. $\sigma^2$ ist ein <u>Hyperparameter</u>).

Wir wählen wieder die Lineare Funktion $f_\Theta(x) = x^T\Theta' + \Theta_0$

**Theorem 13.4**

Sei wieder $\Omega \in \mathbb{R}^{n \times (D+1)}$ die Feature matrix und $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

Der Maximum-Likelihood Schätzer ist

1. nicht eindeutig, falls $r(\Omega) < D+1$.
2. eindeutig bestimmt als
$$\Theta_{MLE} = \Omega^+ Y$$
falls $r(\Omega) = D+1$.

**Beweis**

Es ist $\Theta_{MLE} \in \operatorname{argmax} L(\Theta)$ mit $L(\Theta) = \prod_{i=1}^{n} P_\Theta(y_i | x_i)$.

Wir betrachten
$$\ell(\Theta) = \log L(\Theta) = \sum_{i=1}^{n} \log P_\Theta(y_i | x_i).$$

Anstatt $L(\Theta)$ können wir $\ell(\Theta)$ maximieren.

Es ist
$$\begin{aligned}
\ell(\Theta) &= \sum_{i=1}^{n} \log \Phi(y_i | f_\Theta(x_i), \sigma^2) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - f_\Theta(x_i))^2\right) \\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{n} \left(-\frac{1}{2\sigma^2}(y_i - f_\Theta(x_i))^2\right) \\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f_\Theta(x_i))^2 \\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \|Y - \Omega\Theta\|^2.
\end{aligned}$$

$\leadsto \Theta_{MLE} = \operatorname{argmin}_\Theta \|Y - \Omega\Theta\|^2$. Die Aussage folgt nun wie im Beweis von Thm. 13.2.

$\square$