Vorlesung 14 Lineare & nicht-lineare Regression

Lineare Regression: Wir haben das deterministische Modell

$$y = f_\Theta(x) = x^T\Theta' + \Theta_0, \qquad x \in \mathbb{R}^D, \; y \in \mathbb{R}$$

wobei $\Theta' = (\Theta_1, \dots, \Theta_D)^T \in \mathbb{R}^D$, $\Theta_0 \in \mathbb{R}$.

Wir haben das stat. Modell

$$y \sim N(f_\Theta(x), \sigma^2) \quad \text{mit } \sigma^2 \text{ fest } (= \text{ein Hyperparameter})$$

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

Der MAP ist Θ_{MAP} mit $\Theta_{MAP} = \arg\max P(\Theta \mid X, Y)$.

Wir wählen dazu den Prior $\Theta \sim N(\mu, \Sigma)$, $\mu \in \mathbb{R}^{D+1}$, $\Sigma \in \mathbb{R}^{(D+1)\times(D+1)}_{\text{pos. def.}}$

Theorem 14.1.

Sei $\mu \in \mathbb{R}^{D+1}$, $\Sigma \in \mathbb{R}^{(D+1)\times(D+1)}$ pos. definit. Für den Prior $\Theta \sim N(\mu, \Sigma)$ haben wir:

$$\Theta_{MAP} = (\Omega^T\Omega + \sigma^2\Sigma^{-1})^{-1}(\Omega^T Y + \sigma^2 \Sigma^{-1}\mu).$$

mit

$$\Omega = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}^T \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

und unter der Annahme, dass $\Omega^T\Omega + \sigma^2\Sigma^{-1}$ invertierbar ist.

Beweis

Sei $\alpha(\Theta) = P(\Theta \mid X, Y)$ die Posteriori-Funktion.

$\Theta_{MAP} = \arg\max \alpha(\Theta)$. Es gilt:

$$\alpha(\Theta) = P(\Theta \mid X, Y) = P(\Theta) \frac{P(Y \mid X, \Theta)}{P(Y \mid X)}.$$

nach dem Satz von Bayes. Sei $C := -\log P(Y \mid X)$.

Dann gilt:

$$\log \alpha(\Theta) = \log P(\Theta) + \log P(Y|X,\Theta) + c.$$

und $$P(\Theta) = \frac{1}{\sqrt{(2\pi)^{D+1}\det(\Sigma)}} \exp\left(-\frac{1}{2}(\Theta-\mu)^T \Sigma^{-1}(\Theta-\mu)\right)$$

$$P(Y|X,\Theta) = \prod_{i=1}^{n} P(y_i|x_i,\Theta) = \prod_{i=1}^{n} P_\Theta(y_i|x_i)$$

$$= \frac{1}{\sqrt{2\sigma^2}^n} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - f_\Theta(x_i))^2\right)$$

$$= \frac{1}{\sqrt{2\sigma^2}^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f_\Theta(x_i))^2\right)$$

$$= \frac{1}{\sqrt{2\sigma^2}^n} \exp\left(-\frac{1}{2\sigma^2}\|Y - \Omega\Theta\|^2\right)$$

Wir erhalten:

$$\log \alpha(\Theta) = -\frac{1}{2}(\Theta-\mu)^T\Sigma^{-1}(\Theta-\mu) - \frac{1}{2\sigma^2}\|Y-\Omega\Theta\|^2 + c'$$

mit c' unabh. von Θ.

Um α zu maximieren sehen wir $\nabla_\Theta \log \alpha(\Theta) = 0$.

$$\nabla_\Theta(\log \alpha(\Theta)) = -\Sigma^{-1}(\Theta-\mu) + \frac{1}{\sigma^2}\Omega^T(Y-\Omega\Theta)$$

Mit $\nabla_\Theta(\log\alpha(\Theta)) = 0$ gilt:

$$\Sigma^{-1}(\Theta-\mu) = +\frac{1}{\sigma^2}\Omega^T(Y-\Omega\Theta).$$

$$\Rightarrow \quad \Sigma^{-1}\Theta - \Sigma^{-1}\mu = +\frac{1}{\sigma^2}\Omega^T Y - \frac{1}{\sigma^2}\Omega^T\Omega\Theta$$

$$\Rightarrow \quad \Sigma^{-1}\mu + \frac{1}{\sigma^2}\Omega^T Y = \left(\frac{1}{\sigma^2}\Omega^T\Omega + \Sigma^{-1}\right)\Theta$$

$$\Rightarrow \quad \Theta = (\Omega^T\Omega + \sigma^2\Sigma^{-1})^{-1}(\Omega^T Y + \sigma^2\Sigma^{-1}\mu). \quad \square$$

In den Beweisen zu ERM, RR, MLE und MAP ist das Modell f_Θ in Form der Zeilen von Ω eingegangen.
D.h., wir können einfach das linear Modell zum <u>nicht-linearen</u> Modell erweitern, indem wir folgendes Modell definieren:

$$f_\Theta(x) = \phi(x)^T \Theta \quad \text{mit } \phi: \mathbb{R}^D \to \mathbb{R}^P \text{ nicht-linear.}$$

Die Feature-Matrix ist dann

$$\Omega = [\phi(x_1) \dots \phi(x_n)]^T.$$

z.B.: • mit $\phi(x) = \binom{1}{x} \in \mathbb{R}^{D+1}$ erhalten wir das Lineare Modell

• für $D=1$ und $\phi(x) = (1, x, x^2, \dots, x^d)^T \in \mathbb{R}^{d+1}$ erhalten wir Polynomielle-Regression, da $f_\Theta(x) = \sum_{i=0}^{d} \Theta_i x^i$.

Die Formeln für Θ_{ERM}, Θ_{RR}, Θ_{MLE} und Θ_{MAP} gelten genauso im nicht-linearen Modell.

Jetzt wollen wir den Bayes'schen Ansatz wählen und die Verteilung $P(\Theta | X, Y)$ direkt berechnen.

<u>Theorem 14.2</u>

Mit obigen Annahmen gilt:

$$(\Theta | X, Y) \sim N(m, S),$$

mit

$$S = (\sigma^{-2} \Omega^T \Omega + \Sigma^{-1})^{-1} \text{ und } m = S\left(\frac{1}{\sigma^2} \Omega^T Y + \Sigma^{-1} \mu\right)$$

<u>Beweis</u> Wir berechnen die Dichte $P(\Theta|X,Y)$.

Dazu berechnen wir $\log P(\Theta|X,Y)$.

$$\log P(\Theta|X,Y) = -\frac{1}{2}(\Theta-\mu)^T\Sigma^{-1}(\Theta-\mu) - \frac{1}{2\sigma^2}\|Y-\Omega\Theta\|^2 + c$$

mit c unabh. von Θ.

Wir setzen

$$Q = -(\Theta-\mu)^T\Sigma^{-1}(\Theta-\mu) - \frac{1}{\sigma^2}(Y-\Omega\Theta)^T(Y-\Omega\Theta)$$

so dass $\quad P(\Theta|X,Y) = \exp(\frac{1}{2}Q + c)$

Es gilt:

$$Q = -\Theta\Sigma^{-1}\Theta + 2\mu^T\Sigma^{-1}\Theta - \mu^T\Sigma^{-1}\mu$$
$$-\frac{1}{\sigma^2}(Y^TY - 2Y^T\Omega\Theta + \Theta^T\Omega^T\Omega\Theta)$$

$$= -\Theta\Sigma^{-1}\Theta + 2\mu^T\Sigma^{-1}\Theta + \frac{1}{\sigma^2}2Y^T\Omega\Theta - \frac{1}{\sigma^2}\Theta^T\Omega^T\Omega\Theta + c'$$

mit c' unabh. von Θ. Wir setzen

$$A_1 = \frac{1}{\sigma^2}\Omega^T\Omega + \Sigma^{-1} = S.$$

$$a = \frac{1}{\sigma^2}\Omega^TY + \Sigma^{-1}\mu$$

Damit:
$$Q = \Theta^TA\Theta - 2a^T\Theta + c'$$
$$= (\Theta-b)^TA(\Theta-b) + c''$$

$m = b = A^{-1}a. = S^{-1}a = S^{-1}(\frac{1}{\sigma^2}\Omega^TY + \Sigma^{-1}\mu).$ $\square$

Wir können den Bayes'schen Ansatz verwenden um

- Den Parameter Θ zu samplen, oder
- Den Parameter "auszuintegrieren".

<u>Proposition 14.3</u> Trainingsdaten $X = [x_1, \dots, x_n]^T$ $Y = [y_1, \dots, y_n]^T$

Gegeben seien Daten $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$.
Sei $(x, y) \in \mathbb{R}^D \times \mathbb{R}$ ein weiterer Punkt. Die Verteilung von $(y \mid x, X, Y)$ ist.

$$(y \mid x, X, Y) \sim N(\phi(x)^T m, \ \phi(x)^T S \ \phi(x) + \sigma^2)$$

(für das nicht-lineare Modell bzgl. $\phi(x)$).

<u>Beweisidee</u> Verwende die Formel für Marginal-Density:

$$P(y \mid x, X, Y) = \int_{\mathbb{R}^P} P(y \mid x, X, Y, \Theta) \cdot P(\Theta \mid X, Y) \, d\Theta$$

Die bisherige Diskussion handelt von (nicht)linearen Funktionen

$$f_\Theta : \mathbb{R}^D \to \mathbb{R}.$$

Wir können die Theorie zu Funktion

$$f_\Theta : \mathbb{R}^D \to \mathbb{R}^N$$

erweitern, indem wir

$$(*) \qquad f_\Theta(x) = \begin{bmatrix} \phi(x)^T \Theta_1 \\ \vdots \\ \phi(x)^T \Theta_N \end{bmatrix}$$

mit $\Theta = [\Theta_1, \dots, \Theta_N] \in \mathbb{R}^{P \times N}$
$\phi : \mathbb{R}^D \to \mathbb{R}^P$.

setzen und die Loss Funktion

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

bzw das stat. Modell

$$(y|x) \sim N(f_\theta(x), \sigma^2 \mathbb{1}_N).$$

Weil $\|y-\hat{y}\|^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, können wir jedes θ_i, $i=1,\dots,N$ separat optimieren. D.h. für jeden Eintrag von f_θ in $(*)$ können wir die 1-dim. Schätzer von oben verwenden.