

Sparse Recovery I

"dünn besetzt"

Funktionen $\rightarrow M$ Gewichtungen unabhängige Größen $\in \mathbb{R}^p$
 $\uparrow$ $\uparrow$ $\uparrow$
 Bias # Größen

Lineare Regression:

$\in \mathbb{R}$
abhängige Größe

$$y = \sum_{j=1}^M w_j \cdot \varphi_j(x) + w_0$$

approximierende Funktionen

$$y = w^T \cdot \varphi(x) = \varphi^T(x) \cdot w$$

$$(w_0, w_1, \dots, w_M) \cdot \begin{pmatrix} 1 = \varphi_0(x) \\ \varphi_1(x) \\ \vdots \\ \varphi_M(x) \end{pmatrix}$$

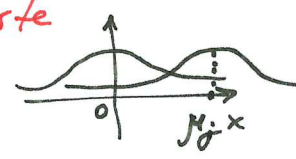
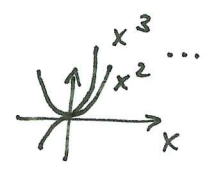
Koordinaten

Beispiele:

• Linear: $\varphi_j(x) = x_j, j=1, \dots, M$
 $p=M$

• polynomiell: $\varphi_j(x) = x^j, j=1, \dots, M$
 $p=1$ Potenzen

• Gaussian: $\varphi_j(x) = e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$
 $p=1$ Mittelwerte Varianzen



Daten:

$$x^i \in \mathbb{R}^p, i=1, \dots, n$$

$$y^i \in \mathbb{R}$$

Datensätze

Wie ist die entsprechende Gewichtung w^* zu bestimmen?

Maximum Likelihood:

$$y^i = \varphi^T(x^i) \cdot w + \varepsilon$$

stochastischer Fehler

$$\Rightarrow y^i \sim \mathcal{N}(\varphi^T(x^i) \cdot w, \sigma^2)$$

normal-verteilt $\sim \mathcal{N}(0, \sigma^2)$

Mittelpunkt = 0
Varianz = σ^2

$$p(y^i; w, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(y^i - \varphi^T(x^i) \cdot w)^2}{2\sigma^2}}$$

Dichte der Normalverteilung

"welche w und σ erklären die Daten am besten?"

$$L(w, \sigma) := \prod_{i=1}^n p(y^i; w, \sigma) = \frac{1}{(\sqrt{2\pi} \cdot \sigma)^n} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \varphi^T(x^i) \cdot w)^2}$$

$\sim$ Bayes a-posteriori mit gleichverteiltem Prior w, σ

i.i.d. Datenmessungen

$$\max_{w, \sigma} \ln L(w, \sigma^2) =$$

$$= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \varphi^T(x^i) \cdot w)^2$$

$$y = \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix}, \quad \Phi(x) = \begin{pmatrix} \varphi^T(x^1) \\ \vdots \\ \varphi^T(x^n) \end{pmatrix}$$

quadratisches Fehler

$$= \|y - \Phi(x) \cdot w\|^2$$

n |

$$\min_{w, \sigma} n \cdot \ln \sigma + \frac{1}{2\sigma^2} \|y - \Phi(x) \cdot w\|^2$$

Optimalitätskriterium (notwendig):

$$0 \stackrel{!}{=} \nabla_w = \frac{1}{2\sigma^2} (-2 \cdot \Phi^T y + 2 \cdot \Phi^T \Phi w)$$

$$\Rightarrow \underbrace{\Phi^T y = \Phi^T \Phi w}_{\text{Normalengleichung}} \Rightarrow w^* = \underbrace{(\Phi^T \Phi)^{-1}}_{\text{falls invertierbar}} \Phi^T y$$

$(y - \Phi \cdot w)^T \cdot (y - \Phi \cdot w) = y^T y - 2 \cdot y^T \Phi \cdot w + w^T \Phi^T \Phi \cdot w$
 (da $w^T \Phi^T y = y^T \Phi \cdot w$)
 Transponieren!

Wann ist $\Phi^T \Phi$ invertierbar?

$$_{M+1} \begin{bmatrix} & n \\ & \end{bmatrix} \begin{bmatrix} & M+1 \\ & \end{bmatrix} = \begin{bmatrix} & M+1 \\ & \end{bmatrix}$$

($n > M+1$)

$\uparrow$ # Daten $\uparrow$ # Funktionen

$\xrightarrow{\text{SVD}}$ alle Eigenwerte von $\Phi^T \Phi$ verschwinden nicht
 $\Leftrightarrow$ Rang von Φ ist $M+1$
 $\Leftrightarrow$

Mindestens $M+1$ Messungen
 (ii) linear polynomiell
 $\Phi^T(x_{i_k}), k=1, \dots, M+1$ sind linear unabhängig

Speziell Bias w_0^* :

$$\begin{pmatrix} 1 & \dots & 1 \\ & * & \end{pmatrix} \cdot \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ & * & \end{pmatrix} \cdot \begin{pmatrix} 1 & \varphi_1(x^1) & \dots & \varphi_M(x^1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \varphi_1(x^n) & \dots & \varphi_M(x^n) \end{pmatrix} \cdot \begin{pmatrix} w_0^* \\ w_1^* \\ \vdots \\ w_M^* \end{pmatrix}$$

$$\sum_{i=1}^n y^i = \sum_{i=1}^n w_0^* + \sum_{i=1}^n \sum_{j=1}^M w_j^* \varphi_j(x^i)$$

$$\Rightarrow w_0^* = \underbrace{\frac{1}{n} \sum_{i=1}^n y^i}_{\text{Mittelwert von abhängiger Größe}} - \sum_{j=1}^M w_j^* \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n \varphi_j(x^i)}_{\text{Mittelwert von Messungen}} \quad - \text{"Mittelwertanpassung"}$$

$$0 \stackrel{!}{=} \nabla_\sigma = \frac{n}{\sigma} - \frac{1}{\sigma^3} \|y - \Phi(x) \cdot w^*\|^2 \Rightarrow \sigma^{*2} = \frac{1}{n} \|y - \Phi(x) \cdot w^*\|^2$$

$\uparrow$ Varianz des Fehlers
 Mittleres Residuum

Optimalitätskriterium (hinreichend):
 bzgl. w

Gradient

$$\nabla_w = \frac{\Phi^T \Phi \cdot w - \Phi^T y}{\sigma^2}$$

Hessematrix

$$H_w = \frac{\Phi^T \Phi}{\sigma^2} \Rightarrow f^T \cdot \frac{\Phi^T \Phi}{\sigma^2} \cdot f = \left(\frac{\Phi \cdot f}{\sigma}\right)^T \cdot \frac{\Phi \cdot f}{\sigma} = \left\| \frac{\Phi \cdot f}{\sigma} \right\|^2 \geq 0 \quad \forall f \in \mathbb{R}^{M+1}$$

Konvexe Funktion

$\hookrightarrow w^*$ - Minimum!

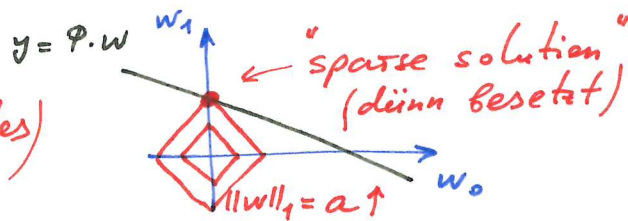
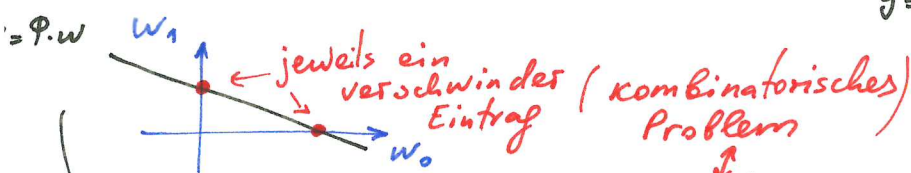
Sparsity: Modell mit möglichst wenigen Parametern
z.B. variable selection in economic analysis

$$\min_w \|w\|_0 \quad \text{s.t.} \quad y_i = \varphi^T(x^i) \cdot w, \quad i=1, \dots, n$$

$$\Leftrightarrow y = \Phi(x)w \quad \text{"lineares Gleichungssystem"}$$

$$= |\{j \mid w_j \neq 0\}| = |S| \quad \text{Support von } w.$$

"0-Norm" $\uparrow$ Anzahl von nichtverschwindenden Einträgen



Approximation wegen Lösbarkeit:

$$P: \min_w \|w\|_1 \quad \text{s.t.} \quad y = \Phi(x) \cdot w$$

$$= \sum_{j=0}^M |w_j| - \text{Summennorm} - \text{konvex als Summe von konvexen Funktionen 1.1}$$

Frage: Lösen s -dünn-besetzte Vektoren mit $\|w\|_0 = s$ die 1-Norm-Relaxation P ?

Dazu: Sei $v \in \mathbb{R}^{M+1}$ mit $y = \Phi(x) \cdot v$

$$u := w - v$$

$$\|w\|_1 = \|w_S\|_1 = \|w_S - v_S + v_S\|_1 \leq \|w_S - v_S\|_1 + \|v_S\|_1 = \|u_S\|_1 + \|v_S\|_1$$

$$(v_S)_j = \begin{cases} v_j, & \text{falls } j \in S \\ 0, & \text{sonst} \end{cases} \quad \text{Support von } w$$

Nullspace (Nullraum von $\Phi(x)$)

$$\Phi(x) \cdot u = \Phi(x) \cdot w - \Phi(x) \cdot v = y - y = 0 \Rightarrow u \in \text{Null}(\Phi(x))$$

Annahme: $\forall u \in \text{Null}(\Phi(x)) \quad \|u_S\|_1 \leq \|u_{S^c}\|_1$ für alle $|S|=s$
(s -null space property) $\nwarrow$ Komplement

$$u = (u_S, u_{S^c})$$

Begründung: $s < M+1 \Rightarrow$

$$\sum_{j \in S} |u_j| \leq \sum_{j \notin S} |u_j|$$

wenige Terme $\quad$ viele Terme

$$\Rightarrow \|w\|_1 \leq \|u_S\|_1 + \|v_S\|_1 = \underbrace{\|w_{S^c} - v_{S^c}\|_1}_{=0} + \|v_S\|_1 = \|v\|_1 \Rightarrow w \text{ löst } P.$$

$$\min_w \|w\|_1 \quad \text{s.t.} \quad \|y - \Phi(x) \cdot w\|^2 \leq \delta \quad (\text{statt } y = \Phi(x) \cdot w)$$

Modellungenauigkeiten bzgl. Daten

$$\text{Lagrange-Ansatz: } \nabla_w [\|w\|_1 - \lambda (\delta - \|y - \Phi(x) \cdot w\|^2)] = 0$$

LASSO:

$$\min_w \|w\|_1 + \lambda \cdot \|y - \Phi(x) \cdot w\|^2 \quad \text{bzgl. } w.$$

$$\lambda := \frac{1}{n}$$

$$\min_w \underbrace{\alpha \cdot \|w\|_1 + \frac{1}{2} \|y - \Phi(x) \cdot w\|^2}_{\text{Regularisierung lineare}}$$