Block 2: Lineares Modell, PCA und Klassifikation

Wintersemester 2018/19

Das lineare Modell

Lineare Regression

Beobachtungen: Paare $(Y_1, x_1), \ldots, (Y_n, x_n) \in \mathbb{R} \times \mathbb{R}^p$, so dass

$$Y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

für nicht beobachtbares (unbekanntes) $\beta \in \mathbb{R}^p$ und i.i.d. Fehler ε_i mit $\mathbb{E}[\varepsilon_i] = 0,\ i = 1, \ldots, n$.

Ziel: Finde ein 'gutes' $\hat{\beta} \approx \beta$ (Schätzer).

Beispiel:

- Y_i Verkaufszahlen von Produkt i,
- $x_i^\top = (x_{i1}, x_{i2}, x_{i3})$ Budgetaufwendungen für Bewerbung von Produkt i im Fernsehen (x_{i1}), Radio (x_{i2}) und in Tageszeitungen (x_{i3}) ('Regressoren').

Einfache und multiple lineare Regression

Ein Spezialfall stellt die einfache lineare Regression dar:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

d.h. $\beta = (\beta_0, \beta_1)^\top$ und $X = (1_{\mathbb{R}^n} \ (x_i)_{1 \leq i \leq n}) \in \mathbb{R}^{n \times 2}$.

Der Parameter β_0 heißt intercept und wird üblicherweise auch bei der multiplen linearen Regression verwendet, so dass

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Lösungsstrategie

Notation: Für $Y = (Y_1, \ldots, Y_n)^\top$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ schreibe

$$Y = X\beta + \varepsilon,$$

wobei $X \in \mathbb{R}^{n \times p}$ ('Designmatrix') aus den Zeilen $x_i^\top$ besteht.

Idee: Minimiere bzgl. der Summe quadrierter 'Residuen'

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - x_i^\top b)^2 = \operatorname{argmin}_{b \in \mathbb{R}^p} \|Y - X\beta\|^2.$$

Explizite Lösung: 'Kleinster-Quadrate-Schätzer'

- nimm an, dass X vollen Spaltenrang ($= p$) hat
- $X\hat{\beta}$ = Bestapproximation von Y im Spaltenraum von $X \Rightarrow$ $X\hat{\beta} = Pi_X Y$
- kann zeigen, dass $Pi_X = X(X^\top X)^{-1} X^\top \Rightarrow \hat{\beta} = (X^\top X)^{-1} X^\top Y$

Orthogonalprojektionen

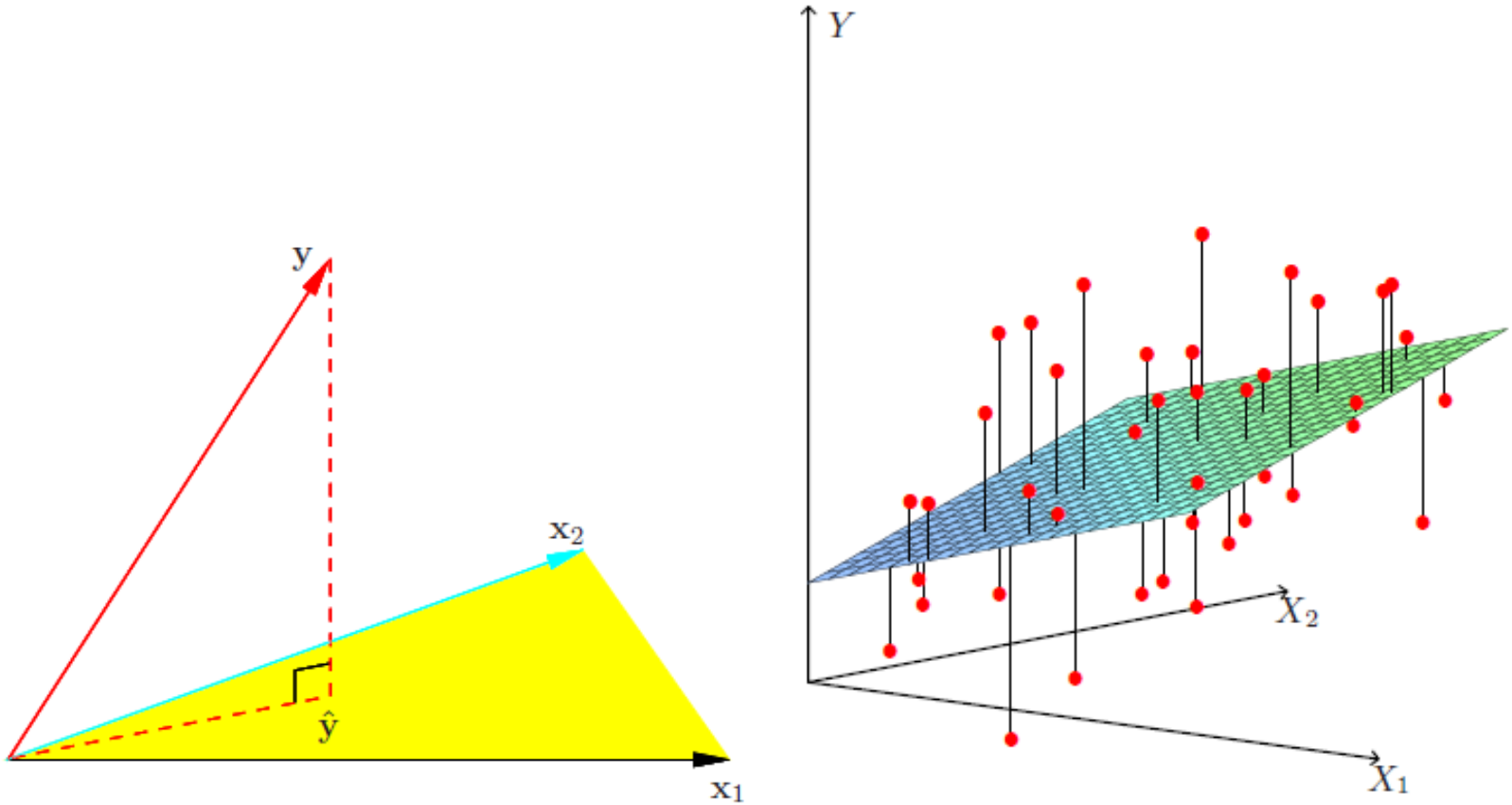

Figure: Projektion der Daten auf den Bildraum (links), Regressionsebene $X\hat{\beta}$ für X variabel (rechts); Bilder aus: *An Introduction to Statistical Learning*, James, Witten, Hastie, Tibshirani, Link)

Test der Koeffizienten

Für $j \in \{1, \ldots, p\}$ fix betrachte das Testproblem

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0.$$

Unter der Annahme $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d. betrachte den t-Test:

$$\varphi(Y) = \mathbb{1}_{\{|T_{0,n-p}(Y)| > q_{t(n-p);1-\alpha/2}\}} \text{ with } T_{0,n-p}(Y) := \frac{(\hat{\beta})_i}{\hat{\sigma}\sqrt{e_i^\top (X^\top X)^{-1} e_i}},$$

where e_i i-th unit vector, $\alpha \in (0, 1)$, $q_{a;\gamma}$ das γ-Quantil der sogenannten t-Verteilung mit a Freiheitsgraden und Varianzschätzer:

$$\hat{\sigma}^2 = \frac{|(I_n - Pi_X)|^2}{n - k}.$$

Klassifikation

Klassifikation

- Gegeben: Daten $(X_1, Y_1), \ldots, (X_n, Y_n)$, $X_k \in \mathbb{R}^p$, $Y_k \in \{0, 1\}$
- Gesucht: 'Klassifizierer' $C : \mathbb{R}^p \to \{0, 1\}$, so dass für 'neue' unabhängige Samples (X, Y) gilt: $C(X) \approx Y$
- diesmal sind X_k und Y_k zufällig
- Beispiel: SPAM-Klassifikation von Emails
  - $1 =$SPAM, $0 =$HAM
  - X_k = Liste aller Worte in Email k
  - Asymmetrie in SPAM vs. HAM
- Training vs. Testen:
  - teile Datensatz in zwei Teilmengen auf
  - auf dem Ersten *trainiere* den Klassifizierer
  - auf dem Zweiten *teste*

0-1-Risiko

- ein "guter" classifier minimiert das *0-1-Risiko*
  $P(C(X) \neq Y) = \mathbb{E}[|\mathbb{1}(C(X) = 1) - \mathbb{1}(Y = 1)|]$
- es gilt:

$$
\begin{aligned}
P(C(X) \neq Y) &= 1 - P\left(C(X) = Y\right) \\
&= 1 - \int P\left(C(x) = Y | X = x\right) P^X\left(dx\right) \\
&= 1 - \int \Big(\mathbb{1}\left(C(x) = 1\right) P\left(Y = 1 | X = x\right) \\
&\qquad + \mathbb{1}\left(C(x) = 0\right) P\left(Y = 0 | X = x\right) \Big) P^X\left(dx\right)
\end{aligned}
$$

$\Rightarrow$ maximiere punktweise den Integranden bezüglich x, d.h. wähle $C(x) = k$, wenn $P(Y = k|X = x)$ maximal

Naive Bayes

- schätze $P(Y = 1|X = x)$ aus den Daten
- Satz von Bayes (für diskrete Zufallsvariablen X und $P(X = x) \neq 0$):

$$P\left(Y = 1|X = x\right) = \frac{P\left(X = x|Y = 1\right) P\left(Y = 1\right)}{P\left(X = x\right)}$$

$$\Rightarrow P(X = x|Y = 1) \approx \frac{\sum_{k=1}^{n} \mathbb{1}_{(X_k=x, Y_k=1)}}{\sum_{k=1}^{n} \mathbb{1}_{(Y_k=1)}}$$

- Beispiel: SPAM-Klassifikation
  - $X_k =$ Liste aller Worte in Email k
  - Problem: X_k sehr hochdimensional

$\Rightarrow$ *naive Bayes*: Wir nehmen an, dass alle Features, bedingt auf Y, unabhängig sind, d.h.

$$P(X = x|Y = 1) = P\left(X^{(1)} = x_1, \ldots, X^{(p)} = x_p|Y = 1\right)$$

$$= \prod_{k=1}^{p} P\left(X^{(k)} = x_k|Y = 1\right)$$

Naive Bayes classifier

- schätze nun $P(X^{(k)} = x_k | Y = 1) \approx \frac{\sum_{j=1}^{n} \mathbb{1}(X_j^{(k)} = x_k, Y_j = 1)}{\sum_{j=1}^{n} \mathbb{1}(Y_j = 1)} := \Phi_{x_k|1}$, ähnlich $\Phi_{x_k|0}$
- für $x \in \mathbb{R}^p$:
  - bestimme $\Gamma_1 = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(Y_j = 1)$, $\Gamma_0 = 1 - \Gamma_1$
  - bestimme $\Phi_{x_k|1}$, $\Phi_{x_k|0}$
  - berechne $\Phi_{x,1} := \prod_{k=1}^{p} \Phi_{x_k|1}$, $\Phi_{x,0} := \Pi_{k=1}^{p} \Phi_{x_k|0}$
  - definiere den classifier

$$C(x) = \mathbb{1}\left(\Phi_{x,1} \cdot \Gamma_1 > \Phi_{x,0} \cdot \Gamma_0\right)$$

- normalerweise verwende stattdessen

$$C(x) = \mathbb{1}\left(\sum_{k=1}^{p} \log\left(\Phi_{lx_k1}\right) + \log \Gamma_1 > \sum_{k=1}^{p} \log\left(\Phi_{lx_k0}\right) + \log \Gamma_0\right)$$

Hauptkomponentenanalyse (PCA)

Hauptkomponentenanalyse

Ziel: Herausarbeiten grundlegener Strukturen bzw. Charakteristika in Datensätzen, welche die Dimension reduzieren und zum Klassifizieren neuer Daten verwendet werden können.

Beispiel:

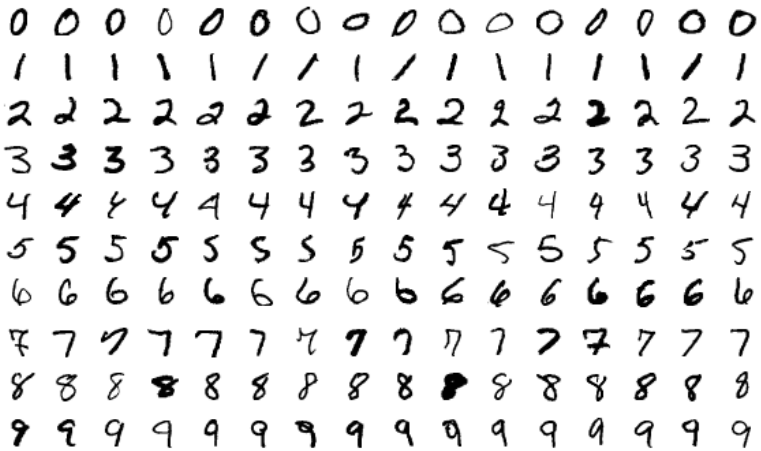

Figure: Handgeschriebene Ziffern aus dem MNIST-Datensatz

Dimensionsreduktion

- Gegeben: p-dimensionale Daten $X_1, \ldots, X_n \in \mathbb{R}^p$, wobei p "groß" & keine Modellannahme (wie z.B. im linearen Modell) gelten
- Ziel: projeziere X_k auf einen affinen Unterraum von viel kleinerer Dimension $q \ll p$
- Idee:
  - finde $\mu \in \mathbb{R}^p$ und Orthogonalprojektion $\Pi \in \mathbb{R}^{p \times p}$ mit Rang q, so dass $\sum_{k=1}^n \|X_k - \mu - \Pi X_k\|^2$ minimal wird
  - oder: finde $f : \mathbb{R}^q \to \mathbb{R}^p$ und $v_1, \ldots, v_n \in \mathbb{R}^q$, so dass $\sum_{k=1}^n \|X_k - f(z_k)\|^2$ minimal wird, wobei $f(x) = Ax + \mu$, $A \in \mathbb{R}^{p \times q}$, $A^\top A = I_q$, $\mu \in \mathbb{R}^p$
- durch Ableiten nach μ und A findet man als Lösungen $\mu = \bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$, $A = (w_1, \ldots, w_q)$, $z_k = A^\top (X_k - \mu)$, wobei $W = (w_1, \ldots, w_p) \in \mathbb{R}^{p \times p}$ und $X^\top X = W \Lambda W^\top$ die Eigenwertzerlegung von $X^\top X$ ist (d.h. $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$)

PCA (= Hauptkomponentenanalyse)

- w_k = Hauptkomponenten von $X^T X$
- es gilt für f oben: $\sum_{k=1}^{n} \|X_k - f(z_k)\|^2 = \sum_{k=q+1}^{n} \lambda_k$
  $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0)$
- Achtung:
  - wenn $p \gg n$ groß, dann ist die Eigenwertzerlegung von $X^\top X \in \mathbb{R}^{p \times p}$ aufwendig
  - $X^T X$ hat dann Rang $\min(n, p) = n$
  - besser: Eigenwertzerlegung von $XX^\top = VDV^\top \in \mathbb{R}^{n \times n}$
  - dann: $XX^\top v_i = d_i v_i$ für $V = (v_1, \ldots, v_n)$ und $D = \operatorname{diag}(d_1, \ldots, d_n)$
    $\Rightarrow \left(X^T X\right) X^\top v_i = d_i X^\top v_i \Rightarrow X^\top v_i$ ist Eigenvektor von $X^\top X$ zum Eigenwert d_i

Vorgehensweise

Algorithmus zum Klassifizieren von X_{n+1} anhand von $X_1, \ldots, X_n$:

- Wähle $q \in \mathbb{N}$ fix.
- Finde die q Hauptkomponenten von bzgl. $X_1, \ldots, X_n$.
- Projiziere $\Pi_q X_1, \ldots, \Pi_q X_n$ und $\Pi_q X_{n+1}$ mittels des durch den Hauptkomponenten erzeugten Unterraum.
- Berechne $X^*_{n+1} := \operatorname{argmin}_{1 \leq i \leq n} \|\Pi_q X_i - \Pi_q X_{n+1}\|$.
- Entscheide (ggf. mittels eines Schwellwertes für obige Distanz), ob X_{n+1} hinreichend gut die Charakteristika von X^*_{n+1} teilt.