BZQ II: Stochastikpraktikum

Block 3: Lineares Modell, Klassifikation, PCA

Randolf Altmeyer

January 9, 2017

Überblick

1. Monte-Carlo-Methoden, Zufallszahlen, statistische Tests
2. Nichtparametrische Methoden
3. **Lineares Modell, Klassifikation, PCA**
4. Markov-Chain-Monte-Carlo-Verfahren
5. Simulation stochastischer Prozesse

Überblick

1. Lineares Modell (lineare Regression)
2. PCA
3. Klassifikation

Literatur:

- Mathias Trabs, Markus Reiß, Moritz Jirak: *Methoden der Statistik* (Skript)

Das lineare Modell (oder auch: Lineare Regression)

- **beobachte** $(x_1, Y_1), \ldots, (x_n, Y_n)$, wobei $x_k \in \mathbb{R}^p$, $Y_k \in \mathbb{R}$ und

$$Y_k = x_k^\top \beta + \varepsilon_k,$$

  $\beta \in \mathbb{R}^p$ unbekannt, ε_k iid und $\mathbb{E}[\varepsilon_k] = 0$
- **Ziel**: finde ein "gutes" $\hat{\beta}$
- **Idee:**

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \left(\sum_{k=1}^{n} \left\| Y_k - x_k^\top b \right\|^2 \right) = \operatorname{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|^2,$$

  wobei $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ hat die $x_k^\top$ als Zeilen
- **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 $\Pi_X = X(X^\top X)^{-1} X^\top \Rightarrow \hat{\beta} = (X^\top X)^{-1} X^\top Y$

Eigenschaften

- Beispiel:
  - Polynomregression: $Y_i = \beta_0 + x\beta_1 + x^2\beta_2 + \cdots + x^{p-1}\beta_{p-1} = (1, x, x^2, \ldots, x^{p-1})(\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$
- **Satz von Gauß-Markov:**
  - $\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[(X^\top X)^{-1} X^\top Y\right] = (X^\top X)^{-1} X^\top \mathbb{E}[Y] = (X^\top X)^{-1} X^\top X\beta = \beta$
  - wenn $\langle v, \beta \rangle$ geschätzt werden soll für $v \in \mathbb{R}^p$, dann ist $\left\langle v, \hat{\beta} \right\rangle$ *best linear unbiased estimator* (BLUE)
- Vorhersagefehler/prediction error:

  $$\mathbb{E}\left[\left(v^\top \hat{\beta} - v^\top \beta\right)^2\right] = \mathrm{Var}\left(v^\top \hat{\beta}\right) = v^\top \mathrm{Var}\left(\hat{\beta}\right) v,$$

  $\mathrm{Var}(\hat{\beta}) = \mathbb{E}[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^\top]$, Bias ist null!
- aber:
  - $\hat{\beta}$ ist nur dann wohldefiniert, wenn X vollen Spaltenrang hat, d.h. wenn $p \leq n$
  - was tun, wenn $p \gg n$?

Regularisierung

- zurück zur Polynomregressin:
  wenn Modellpolynom zu hohen Grad hat, werden die Koeffizienten β_k sehr groß
  $\Rightarrow$ Regularisierung
- l^2-norm (*Ridge-regression*):

$$\hat{\beta}^{RR} = \text{argmin}_{b \in \mathbb{R}^p} \left(\|Y - Xb\|^2 + \lambda \|b\|_{l^2}^2 \right)$$

  $\Rightarrow$ kleinere Koeffizienten werden bevorzugt
- l^0-norm (*subset selection*):

$$\hat{\beta}^{SS} = \text{argmin}_{b \in \mathbb{R}^p} \left(\|Y - Xb\|^2 + \lambda \|b\|_{l^0} \right)$$

  $\Rightarrow$ Koeffizienten werden oft auf 0 gesetzt ($\Rightarrow$ Variablenselektion!)
- l^1-norm (*lasso = least absolute selection and shrinkage operator*):

$$\hat{\beta}^{L} = \text{argmin}_{b \in \mathbb{R}^p} \left(\|Y - Xb\|^2 + \lambda \|b\|_{l^1} \right)$$

  $\Rightarrow$ kleinere Koeffizienten werden bevorzugt *und* Koeffizienten werden oft auf 0 gesetzt

PCA

- **gegeben:** $X_1, \ldots, X_n \in \mathbb{R}^p$, p "groß", *keine* Modellannahme wie im linearen Modell
- **Ziel:** projeziere X_k auf 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 $\sum_{k=1}^n \|X_k - f(v_k)\|^2$ 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 $\mu = \bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$, $A = (w_1, \ldots, w_q)$, $v_k = A^\top (X_k - \mu)$, 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: $\sum_{k=1}^{n} \|X_k - f(v_k)\|^2 = \sum_{k=q+1}^{n} \lambda_k$
  $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0)$
- **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 $D = \mathrm{diag}(d_1, \ldots, d_n)$
    $\Rightarrow \left(X^T X\right) X^\top v_i = d_i X^\top v_i \Rightarrow X^\top v_i$ ist Eigenvektor von $X^\top X$ zum Eigenwert d_i

Klassifikation

- **gegeben:** Daten $(X_1, Y_1), \ldots, (X_n, Y_n)$, $X_k \in \mathbb{R}^p$, $Y_k \in \{0, 1\}$
- **gesucht:** "Klassifikator" $C : \mathbb{R}^p \to \{0, 1\}$, so dass für "neue" unabhängige Samples (X, Y) $C(X) \approx Y$
- 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 classifier
  - auf dem Zweiten *teste*

0-1-Risiko

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

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

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

Naive Bayes

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

$$\mathbb{P}(Y = 1|X = x) = \frac{\mathbb{P}(X = x|Y = 1)\,\mathbb{P}(Y = 1)}{\mathbb{P}(X = x)}$$

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

- 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.

$$\begin{aligned}\mathbb{P}(X = x|Y = 1) &= \mathbb{P}\left(X^{(1)} = x_1, \ldots, X^{(p)} = x_p \middle| Y = 1\right) \\ &= \prod_{k=1}^{p} \mathbb{P}\left(X^{(k)} = x_k \middle| Y = 1\right)\end{aligned}$$

Naive Bayes classifier

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

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

- normalerweise verwende stattdessen

$$C(x) = \mathbf{1}\left(\sum_{k=1}^{p} \log\left(\Phi_{x_k|1}\right) + \log \Gamma_1 > \sum_{k=1}^{p} \log\left(\Phi_{x_k|0}\right) + \log \Gamma_0\right)$$