Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

1. Übungsblatt

1. Betrachte eine mathematische Stichprobe $X_1, \dots, X_n$, wobei $X_1 \sim U([a,b])$ mit unbekannten Parametern $-\infty < a < b < \infty$.

   (a) Formalisiere das statistische Modell.

   (b) Bestimme den MLE $\widehat{\vartheta} = (\widehat{\vartheta}_1, \widehat{\vartheta}_2)$ für den Parameter $(a,b) \in \mathbb{R}^2$.

   (c) Berechne den MSE von $\widehat{\vartheta}_1$ bezüglich a und $\widehat{\vartheta}_2$ bezüglich b.

2. In einem Krankenhaus soll zur Hebammenplanung mit 95% Sicherheit eine Obergrenze für die Verteilung der Geburtenzahl pro Tag angegeben werden. Bekannt sind die Geburtenzahlen $N_1, \dots, N_n$ der vergangenen n Tage.

   (a) Warum können die $N_1, \dots, N_n$ näherungsweise als unabhängig und Poiss(λ)-verteilt angesehen werden mit unbekanntem Parameter $\lambda > 0$?

   (b) Formalisiere das statistische Modell.

   (c) Prüfe das aritmethische Mittel $\widehat{\lambda}$ der $N_1, \dots, N_n$ auf Erwartungstreue und Konsistenz. Berechne den MSE von $\widehat{\lambda}$. Ist $\widehat{\lambda}$ ein MLE?

3*. Lese Kapitel 2.3 "Lab:Introduction to R" aus "Introduction to Statistical Learning". Die dort verwendete Datei "Auto.data" findet man unter www-bcf.usc.edu/∼gareth/ISL/Auto.data.

4. Betrachte eine mathematische Stichprobe $X_1, \dots, X_n$, wobei $X_1 \sim \mathbb{P}_\vartheta$, $\vartheta \in \Theta$. Für $r > 0$ sei $\psi_r : \mathbb{R} \to \mathbb{R}$ die Funktion mit

$$\psi_r(x) = \begin{cases} -r, & x < -r, \\ x, & |x| \le r, \\ r, & x > r. \end{cases}$$

   Sei M_r die Menge der Nullstellen der Funktion

$$\Psi_r : \mathbb{R} \to \mathbb{R}, y \mapsto \Psi_r(y) := \frac{1}{n} \sum_{i=1}^{n} \psi_r(X_i - y), \ y \in \mathbb{R}.$$

   Definiere die Zufallsvariable $\widehat{m}_r = \inf M_r$.

   (a) Zeige, dass $\widehat{m}_r$ für alle $r > 0$ existiert und dass $\Psi_r(\widehat{m}_r) = 0$ gilt.

(b) Gib die Definition eines Medians von $\mathbb{P}_\vartheta$ und von $X_1, \ldots, X_n$ an. Was ist der Unterschied zwischen Median und Erwartungswert von $\mathbb{P}_\vartheta$ bzw. von Median und arithemetischen Mittel von $X_1, \ldots, X_n$?

(c) Zeige, dass $\widehat{m}_r$ fast sicher gegen einen Median (gegen das arithmetische Mittel) von $X_1, \ldots, X_n$ konvergiert für $r \to 0$ ($r \to \infty$).

(d*) Angenommen $\mathbb{P}_\vartheta$ hat einen eindeutigen Median. Zeige, dass der Median von $X_1, \ldots, X_n$ ein konsistenter Schätzer des Medians von $\mathbb{P}_\vartheta$ ist.

(e) Simuliere $n = 200$ unabhängige Zufallsvariablen mit den Verteilungen

i. Lognormalverteilung $\log N(\mu, \sigma)$ mit Parametern $\mu = 1$, $\sigma^2 = 1$, sowie $\mu = 1$, $\sigma^2 = 3$.

ii. Cauchy-Verteilung $\mathrm{Cau}(x_0, \gamma)$ mit Parametern $x_0 = 0$, $\gamma = 1$, sowie $x_0 = 0$, $\gamma = 5$.

Schätze in allen vier Situationen mit R den Median und den Erwartungswert der Verteilungen. Berechne $\widehat{m}_r$ für $r \in \{1/10, 1/2, 1, 2, 5, 10, 20, 50\}$ und stelle die Ergebnisse graphisch dar. Was können wir daraus folgern für die Beziehung zwischen $\widehat{m}_r$, Median und arithmetischem Mittel von $X_1, \ldots, X_n$ folgern?

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 25.10.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

2. Übungsblatt

1. Die Beta-Verteilung Beta(a,b) für $a,b>0$ ist gegeben durch die Dichte

$$f_{a,b}(x) = \mathrm{B}(a,b)^{-1}x^{a-1}(1-x)^{b-1}, \;\; x\in(0,1),$$

wobei $\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ die Beta-Funktion und $\Gamma(a)$ die Gamma-Funktion bezeichnet.

   (a) Skizziere (z.B. mit Hilfe von R) $f_{a,b}$ für $(a,b)\in\{0.5,1,10\}^2$.

   (b) Zeige, dass die Gleichverteilung auf $[0,1]$ eine Beta-Verteilung ist.

   (c) Zeige: Beta(a,b) hat Erwartungswert $\frac{a}{a+b}$ und Varianz $\frac{ab}{(a+b)^2(a+b+1)}$.

2. Betrachte die Beobachtung $X\sim\mathrm{Bin}(n,p)$ für $p\in[0,1]$. Bestimme einen MSE-minimax-Schätzer von p und zeige, dass der MLE nicht MSE-minimax ist. Verwende dazu die folgenden Schritte:

   (a) Betrachte die a-priori-Verteilung $p\sim\mathrm{Beta}(a,b)$ für $a,b>0$. Bestimme die a-posteriori-Verteilung.

   (b) Zeige, dass der Bayes-optimale Schätzer (bzgl. MSE und Beta(a,b)) gegeben ist durch $\widehat{p}_{a,b}=\frac{X+a}{a+b+n}$.

   (c) Bestimme den MSE $R(\widehat{p}_{a,b},p)$ von $\widehat{p}_{a,b}$ und finde $a^*,b^*>0$, so dass $p\mapsto R(\widehat{p}_{a^*,b^*},p)$ konstant ist.

   (d) Zeige, dass aus konstantem MSE und Bayes-Optimalität bereits folgt, dass $\widehat{p}_{a^*,b^*}$ MSE-minimax-Schätzer ist. Folgere, dass der MLE nicht minimax ist.

3. Sei $X_1,\dots,X_n$ eine mathematische Stichprobe, wobei $X_1\sim\mathrm{Exp}(\lambda)$ exponentialverteilt ist mit Parameter $\lambda>0$ und Lebesgue-Dichte

$$f_\lambda(x)=\lambda e^{-\lambda x}, \;\; x\geq 0.$$

Bestimme für $\lambda>0$ eine asymptotisch korrekte Konfidenzmenge zum Niveau $1-\alpha$ für $\alpha\in(0,1)$. Verwende dazu die folgenden Schritte:

   (a) Bestimme den Erwartungswert von X_1.

(b) Beweise: Seien $Y_1, Y_2, \ldots$ reelle Zufallsvariablen, so dass

$$a_n(Y_n - b) \xrightarrow{d} N(0, \sigma^2), \quad n \to \infty,$$

für $b \in \mathbb{R}$, $\sigma^2 > 0$ und eine Folge positiver Zahlen $(a_n)_{n \geq 1}$ mit $a_n \to \infty$. Wenn $g : \mathbb{R} \to \mathbb{R}$ differenzierbar ist bei b und $g'(b) > 0$, dann ist

$$a_n(g(Y_n) - g(b)) \xrightarrow{d} N(0, (g'(b))^2 \sigma^2).$$

Hinweis: Verwende eine Taylorapproximation von g und Slutzky's Lemma.

(c) Verwende (b) für eine geeignete Funktion g, um die Konfidenzmenge zu bestimmen.

4. Simuliere unabhängige Zufallsvariablen $X_1, \ldots, X_n$ für $n \in \{5, 20, 100\}$ mit den Verteilungen

   (a) Ber(p) für $p = 0.1$, $p = 1/2$,

   (b) Exp(λ) für $\lambda = 0.1$, $\lambda = 2$,

   Bestimme Konfidenzintervalle zum Niveau 0.95 für den jeweils unbekannten Parameter mittels Normalapproximation und Bootstrap. Vergleiche die Ergebnisse.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 01.11.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

3. Übungsblatt

1. Eine Münze wird n Mal unabhängig geworfen. Wir wollen zum Niveau $\alpha = 0.05$ testen, ob die Münze fair ist (d.h. $p = 0.5$) oder nicht (d.h. $p \neq 0.5$).

   (a) Gib das statistische Modell, die Nullhypothese und die Alternative an.

   (b) Zeige, dass

   $$\varphi_\alpha(x) = \mathbf{1}\big(\bar{x}_n^{n\bar{x}_n}(1-\bar{x}_n)^{n(1-\bar{x}_n)} > c_\alpha\big) + \gamma_\alpha \mathbf{1}\big(\bar{x}_n^{n\bar{x}_n}(1-\bar{x}_n)^{n(1-\bar{x}_n)} = c_\alpha\big)$$

   ein Likelihood-Quotienten-Test ist für $x \in \{0,1\}^n$, $c_\alpha \geq 0$, $\gamma_\alpha \in [0,1]$ und $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$.

   (c) Bestimme c_α und γ_α für $n = 6$, so dass φ_α exakt das Niveau α erreicht.

2. Im Jahr 2015 wurden in Berlin 38030 Kinder geboren, davon waren 19614 Jungen. Die Wahrscheinlichkeit einer Jungengeburt sei p.

   (a) Verwende die 1. Aufgabe und Normalapproximation, um einen Test (mit kritischen Werten) zum Niveau $\alpha = 0.05$ von $H_0 : p = 0.5$ gegen $H_0 : p \neq 0.5$ zu bestimmen. Kann die Nullhypothese abgelehnt werden?
   Hinweis: Beachte, dass die Funktion $y \mapsto y^y(n-y)^{n-y}$ für $y \in [0,n]$ symmetrisch um $n/2$ ist und für $y \geq n/2$ wächst.

   (b) Bestimme den p-Wert.

3. Beweise den Satz aus der Vorlesung über gleichmäßig beste Tests bei monotonen Likelihood-Quotienten.

4. Lese den Artikel zur Schokoladen-Studie (den man hier findet).

   (a) Erkläre was mit der Studie nicht stimmt mit Hilfe der folgenden Konstruktion: Sei $(\varphi_i)_{i\in\mathbb{N}}$ eine Familie unverfälschter und unabhängiger Tests (d.h. die Familie der Zufallsvariablen $(\varphi_i)_{i\in\mathbb{N}}$ ist unabhängig), die das Niveau $\alpha \in (0,1)$ exakt erreichen, alle definiert auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Zeige, dass $\mathbb{P}(\bigcup_{i=1}^n \{\varphi_i = 1\}) \to 1$ für $n \to \infty$.

   (b) Schlage den Begriff "Bonferroni-Korrektur" nach und erkläre wie diese das Problem in (a) löst.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 08.11.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

4. Übungsblatt

1. Betrachte ein gewöhnliches lineares Modell $Y = X\beta + \varepsilon$ mit Fehlern $\varepsilon_i \sim N(0, \sigma^2)$ für unbekanntes $\sigma > 0$. Bestimme einen Likelihood-Quotienten-Test von $H_0 : \beta = \beta_0$ gegen $H_1 : \beta \neq \beta_0$ zum Niveau $\alpha \in (0, 1)$.

2. Betrachte unabhängige Messwerte $Y_1, \dots, Y_n$ mit $\mathbb{E}[Y_i] = \mu \in \mathbb{R}$, $\mathrm{Var}(Y_i) = \sigma_i^2$ für bekannte $\sigma_i > 0$ und $\mathrm{Cov}(Y_i, Y_j) = 0$ für $i \neq j$.

   (a) Schreibe dies als lineares Modell und berechne den gewichteten Kleinste-Quadrate-Schätzer $\widehat{\mu}$ für μ.

   (b) Vergleiche die Varianzen von $\widehat{\mu}$ und dem Stichprobenmittel $\bar{Y}$. Zeige, dass $\widehat{\mu}$ die kleinste Varianz unter allen linearen, erwartungstreuen Schätzern besitzt.

3. Betrachte Beobachtungen $X_0, X_1, \dots, X_n$, die einem *autoregressiven Modell* folgen, d.h. $X_k = \gamma X_{k-1} + \varepsilon_k$, $k = 1, \dots, n$, für unbekanntes $\gamma \in \mathbb{R}$ und i.i.d. Fehler ε_i mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ für bekanntes $\sigma > 0$. Bestimme einen Schätzer mit Hilfe der Methode der kleinsten Quadrate. Ist dieser Schätzer erwartungstreu?

4. Betrachte die einfache lineare Regression $Y_i = ax_i + b + \varepsilon_i$, $i = 1, \dots, n$ für unbekannte $a, b \in \mathbb{R}$ und mit Fehlern $\varepsilon_i \sim \mathrm{Exp}(\lambda)$ für bekanntes $\lambda > 0$. Bestimme den Maximum-Likelihood-Schätzer von a, b.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 15.11.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

5. Übungsblatt

1. Betrachte ein gewöhnliches lineares Modell, wobei die erste Spalte der Matrix X gleich $(1,\dots,1)^\top$ ist. Sei $\bar{Y}$ das arithmetische Mittel der Beobachtungen und sei $\widehat{Y} = X\widehat{\beta}$ für den KQS $\widehat{\beta}$. Dann heißt

$$R^2 := \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

   Bestimmtheitsmaß. Zeige, dass $R^2 \in [0,1]$. Was bedeuten die Fälle $R^2 = 0$ bzw. $R^2 = 1$ statistisch (mit mathematischer Begründung)?

2. Betrachte eine mathematische Stichprobe $(Y_1, X_1), \dots, (Y_n, X_n)$ mit Zufallsvariablen $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$. Wir nehmen an, dass $\mathbb{E}[X_1] = 0$ und dass die Kovarianzmatrix $\mathbb{E}[X_1 X_1^\top] = \Sigma_X \in \mathbb{R}^{p\times p}$ existiert und positiv definit ist. Wir nehmen außerdem an, dass $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$ und $\mathrm{Var}(Y_i|X_i) = \sigma^2$ für deterministische $\beta \in \mathbb{R}^p$, $\sigma > 0$. (Y_i, X_i) folgt dann einem linearen Modell mit zufälligem Design $Y_i = X_i^\top \beta + \varepsilon_i$ für $\varepsilon_i = Y_i - X_i^\top \beta$.

   (a) Bestimme den KQS $\widehat{\beta}$. Bestimme Erwartungswert und Varianz.

   (b) Zeige, dass $\widehat{\beta}$ konsistent ist.

3. Betrachte ein gewöhnliches lineares Modell unter der Normalverteilungsannahme $\varepsilon \sim N(0, \sigma^2 E_n)$. Formuliere die folgenden Situationen als lineare Testprobleme mit unbekanntem $\beta \in \mathbb{R}^p$, konstruiere die zugehörigen Fisher-Statistiken, bestimme deren Verteilung, gib die zugehörigen F-Tests zum Niveau $\alpha \in (0,1)$ und Konfidenzbereiche zum Niveau $1-\alpha$ an:

   (a) $H_0 : \langle \beta, v\rangle = \langle \beta^*, v\rangle$ gegen $H_1 : \langle \beta, v\rangle \neq \langle \beta^*, v\rangle$ für bekanntes $v \in \mathbb{R}^p$ und unbekanntes $\sigma > 0$.

   (b) $H_0 : \beta_j = 0$ gegen $H_1 : \beta_j \neq 0$ für festes $j \in \{1,\dots,p\}$ und unbekanntes $\sigma > 0$.

   (c) (kein F-Test) H_0, H_1 wie in (a), aber für bekanntes $\sigma > 0$.

4. Betrachte zwei unabhängige mathematische Stichproben $X_1,\dots,X_n \sim N(\mu_1, \sigma^2)$, $Y_1,\dots,Y_n \sim N(\mu_2,\sigma^2)$ für unbekannte $\mu_1, \mu_2 \in \mathbb{R}$, aber bekanntes $\sigma > 0$. Bestimme einen Test zum Niveau $\alpha \in 0.05$ von $H_0 : \mu_1 = \mu_2$ gegen $H_1 : |\mu_1 - \mu_2| > \Delta$ für ein vorgegebenes $\Delta > 0$. Wie groß muss n sein, damit der Test gleichmäßig über alle Parameter unter H_1 eine Power von mehr als 90% hat?

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 22.11.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

6. Übungsblatt

1. Untersuche die Paneldaten von der Webseite im Hinblick auf Unterschiede im Nettoeinkommen von Männern und Frauen in den neuen und alten Bundesländern. Betrachte dafür folgende Schritte:

   (a) Führe eine einfache deskriptive Analyse durch (Mittelwerte, Boxplots, gibt es Ausreißer?) und erkläre kurz deine ersten Erkenntnisse.

   (b) Führe eine lineare Regression durch von allen nicht-kategoriellen Variablen auf das Nettoeinkommen („inc_pos“), zuerst für alle Daten, dann getrennt für Frauen bzw. Männer in den alten/neuen Bundesländern. Diskutiere, ob die Annahmen des gewöhnlichen linearen Modells mit normalverteilten Fehlern erfüllt sind (z.B. plotte Residuen bzw. einen QQ-Plot der Residuen, betrachte Hypothesentests, plotte Vorhersagefehler gegen Residuen etc.). Erkläre welche Variablen aus dem Modell entfernt werden können (z.B. mit Hilfe von R^2 oder mit Hilfe des *variance inflation factor*).

   (c) Führe Varianzanalysen durch (zum Niveau $\alpha = 0.05$) für die folgenden Nullhypothesen und diskutiere die Ergebnisse:

      i. das Nettoeinkommen von Männern und Frauen ist gleich,

      ii. das Nettoeinkommen von Frauen in den neuen Bundersländern ist gleich,

      iii. der Bildungsgrad von Frauen in Berlin und im Saarland ist gleich.

   (d*) Gibt es noch andere interessante Fragestellungen, die sich statistisch nachprüfen lassen?

   Hinweise zu R: eventuell hilfreich sind Konzepte/Methoden wie *data frame*, *factor*, *aggregrate*, *aov*, *gather*.

2. Im Modell der einfaktoriellen Kovarianzanalyse (ANCOVA, analysis of covariance) beobachtet man

$$Y_{ij} = \mu_i + \kappa x_{ij} + \varepsilon_{ij}, \quad j = 1, \dots, n_i,\ i = 1, \dots, p,$$

   mit i.i.d Fehlern $\varepsilon_{ij} \sim N(0, \sigma^2)$ und bekannten Kovariaten $x_{ij} \in \mathbb{R}$, während $\sigma > 0$, die Gruppenmittelwerte $\mu_1, \dots, \mu_p \in \mathbb{R}$ und der Regressionskoeffizient $\kappa \in \mathbb{R}$ unbekannt sind.

(a) Gib ein passendes lineares Modell für diese Beobachtungen an. Finde notwendige und hinreichende Bedinungen an $(x_{ij}) \in \mathbb{R}^{p\times n}$ (mit $n = n_1 + \cdots + n_p$), damit die Designmatrix vollen Rang hat.

(b) Bestimme den Kleinste-Quadrate-Schätzer $\widehat{\beta} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p, \widehat{\kappa})^\top$.

(c) Zeige:

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\kappa}x_{ij} - (\bar{Y}_{i\bullet} - \widehat{\kappa}\bar{x}_{i\bullet}))^2 + \sum_{i=1}^{p} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \widehat{\kappa}^2\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})^2.$$

(d) Bestimme die Fisher-Statistik für einen Test der Nullhypothese $H_0 : \mu_1 = \cdots = \mu_p$.

3. Diskutiere, ob die folgenden Verteilungen Exponentialfamilien bilden. Bestimme gegebenenfalls die Funktionen T, η, c, ζ und den natürlichen Parameterraum.

(a) Multinomialverteilung $(M(p_0, \ldots, p_m; n))_{0<p_i<1, \sum_{i=0}^m p_i = 1}$,

(b) p-dimensionale Normalverteilung $(N(\mu, \Sigma))_{\mu\in\mathbb{R}^p}$ mit bekannter Kovarianzmatrix $\Sigma \in \mathbb{R}^{p\times p}$,

(c) Gleichverteilung $(U([0, \vartheta]))_{\vartheta>0}$,

(d) Gammverteilung $(\Gamma(a, b))_{a,b>0}$,

(e) Verteilung des Vektors $Y = (Y_1, \ldots, Y_n)^\top$ aus Aufgabe 4 von Blatt 4.

Abgabe in Zweier- oder Dreier-Gruppen: Aufgabe 1 vor der Vorlesung am Freitag 6.12.19. (Code an altmeyrx@math.hu-berlin.de), übrige Aufgaben vor der Vorlesung am Freitag, 29.11.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

7. Übungsblatt

1. Betrachte eine mathematische Stichprobe $Y_1, \dots, Y_n$, wobei die Verteilung von Y_1, $(\mathbb{P}_\eta^{Y_1})_{\eta \in H}$, einer natürlichen exponentiellen Familie in Dimension $k = 1$ folgt mit Dichte

$$p_\vartheta(x) = c(x) \exp(\eta(\vartheta) T(x) - \zeta(\vartheta)), \quad x \in \mathfrak{X}.$$

   Finde einen asymptotisch korrekten Test von $H_0 : \eta = \eta_0$ gegen $H_1 : \eta = \eta_1$ für $\eta_1 \neq \eta_0$ zum Niveau $\alpha \in (0,1)$.

2. Betrachte das Modell der logistischen Regression mit $p_i = \frac{\exp(\langle x_i, \beta \rangle)}{1+\exp(\langle x_i, \beta \rangle)}$ für gegebene Designpunkte $x_1, \dots, x_n \in \mathbb{R}^2$ und Parameter $\beta \in \mathbb{R}^2$. Für $x \in \mathbb{R}^2$ definiere die Vorhersage $\widehat{p}(x) := \frac{\exp(\langle x, \widehat{\beta} \rangle)}{1+\exp(\langle x, \widehat{\beta} \rangle)}$, wobei $\widehat{\beta} \in \mathbb{R}^2$ der MLE von β ist (wenn er existiert), und die Klassifikation $\widehat{k}(x) := \mathbf{1}(\widehat{p}(x) > 1/2)$.

   (a) Bestimme die geometrische Form der Klassifikationsgrenze $\widehat{B} := \{x \in \mathbb{R}^2 : \widehat{p}(x) = 1/2\}$.

   (b) Wie sieht die Iteration in Fishers Scoring-Methode aus?

   (c) Simuliere i.i.d. Zufallsvariablen $X_1, \dots, X_n$ mit $X_1 \sim N(0, E_2)$ für $n = 100$ und definiere $x_i := X_i$ (zufälliges Design). Erzeuge damit Beobachtungen $Y_1, \dots, Y_n$ der logistischen Regression mit p_i wie oben und $\beta = (2,1)^\top$. Stelle die Daten, die Funktion $\widehat{p} : \mathbb{R}^2 \to [0,1]$, sowie die Klassifikationsgrenze $\widehat{B}$ gemeinsam in einem Koordinatensystem dar.

3. Betrachte das Modell der Poissonregression mit $\log \lambda_i = \eta_i = a x_i + b$ für gegebene Designpunkte $x_1, \dots, x_n \in \mathbb{R}$.

   (a) Diskutiere die Frage, ob ein MLE existiert und ob er bei Existenz eindeutig ist.

   (b) Wie sieht die Iteration in Fishers Scoring-Methode aus?

4. Betrachte ein gewöhnliches lineares Modell mit Fehlern $\varepsilon \sim N(0, \sigma^2 E_n)$.

   (a) Berechne für unbekannte $\beta \in \mathbb{R}^p, \sigma > 0$ den MLE von σ^2.

   (b) Berechne Bias und Varianz des MLEs und des Schätzers $\widehat{\sigma^2} = \frac{\|Y - X\widehat{\beta}\|^2}{n-p}$, wobei $\widehat{\beta}$ der Kleinste-Quadrate-Schätzer für β ist.

(c) Bestimme für bekanntes β und unbekanntes $\sigma > 0$ einen Likelihood-Quotienten-Test für das einseitige Testproblem $H_0 : \sigma^2 = \sigma_0^2$ gegen $H_1 : \sigma^2 > \sigma_0^2$ zum Niveau $\alpha \in (0, 1)$.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 06.12.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

8. Übungsblatt

1. Zeige für die Kullback-Leibler-Divergenz folgende Eigenschaften:

   (a) $KL(\mathbb{P}|\mathbb{Q}) \geq 0$,

   (b) $KL(\mathbb{P}|\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$,

   (c) im Allgemeinen ist $KL(\mathbb{P}|\mathbb{Q}) \neq KL(\mathbb{Q}|\mathbb{P})$ (gib zwei Beispiele an!),

   (d) $KL(\mathbb{P}_\vartheta^{\otimes n}|\mathbb{P}_{\vartheta_0}^{\otimes n}) = nKL(\mathbb{P}_\vartheta|\mathbb{P}_{\vartheta_0})$.

2. Betrachte eine mathematische Stichprobe $X_1, \dots, X_n$ mit $X_1 \sim \mathbb{P}_\vartheta$ für $\vartheta \in \Theta$ und nimm an, dass $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta'}$ für alle $\vartheta, \vartheta' \in \Theta$. Sei $\ell(\vartheta, x) := \log L(\vartheta, x)$ die log-Likelihood von $\mathbb{P}_\vartheta$ bezüglich einem dominierenden Maß ν. Nimm außerdem an, dass $KL(\mathbb{P}_\vartheta|\mathbb{P}_{\vartheta'}) < \infty$ und $KL(\mathbb{P}_\vartheta|\nu) < \infty$ für alle $\vartheta, \vartheta' \in \Theta$.

   (a) Zeige für $\vartheta_0 \in \Theta$ und $n \to \infty$:

   $$-\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, X_i) \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} KL(\mathbb{P}_{\vartheta_0}|\mathbb{P}_\vartheta) - KL(\mathbb{P}_{\vartheta_0}|\nu).$$

   (b) Diskutiere, welcher Parameter in Θ der natürliche Grenzwert des MLEs ist in einem missspezifizierten Modell, so dass $\vartheta_0 \notin \Theta$ für das wahre ϑ_0.

3. Beweise den zweiten Teil des Satzes über die Orakelungleichung (mit der Notation aus dem Satz):

   $$\mathbb{E}[\|\widehat{\mu}^{(\widehat{k})} - \mu\|^2] \leq C_{\alpha,\gamma}(\min_{k=1,\dots,K}(\|(E_n - \Pi^{(k)})\mu\|^2 + \mathrm{Pen}(d_k)) + 2\sigma^2\varepsilon_{\alpha,0}).$$

   Hinweis: Zeige zuerst $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)dt$ für eine Zufallsvariable X mit Werten in $[0, \infty)$.

4. Sei $f : [0,1] \to \mathbb{R}$ eine unbekannte Funktion gegeben durch $f = \sum_{j=1}^m a_j\varphi_j$ für $m \leq n$ mit Koeffizienten $a_j \in \mathbb{R}$ und Funktionen $\varphi_j \in L^2([0,1])$. Definiere das empirische Skalarprodukt $\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(\frac{i}{n})g(\frac{i}{n})$ mit empirischer Norm $\|f\|_n := \langle f, f\rangle_n^{1/2}$ und nimm an, dass die φ_j eine Orthonormalbasis bilden bezüglich $\langle \bullet, \bullet\rangle_n$. Betrachte das Regressionsmodell $Y_i = f(\frac{i}{n}) + \varepsilon_i$ mit i.i.d. Fehlern $\varepsilon_i \sim N(0, \sigma^2)$ für bekanntes $\sigma > 0$.

   (a) Zeige, dass $\widehat{a}_j := \langle Y, \varphi_j\rangle_n := \frac{1}{n}\sum_{i=1}^n Y(i)\varphi_j(\frac{i}{n})$ MLE von a_j ist für $j = 1, \dots, n$.

(b) Für Unterräume $S_k := \text{span}(\varphi_1, \ldots, \varphi_k)$ formuliere approximierende lineare Modelle $X^{(k)}\beta^{(k)}$ und bestimme die entsprechenden Kleinste-Quadrate-Schätzer $\widehat{\beta}^{(k)}$.

(c) Formuliere eine Orakelungleichung wie in Aufgabe 3 für $\mathbb{E}[\|\widehat{\mu}^{(\widehat{k})} - \mu\|_n^2]$ mit $\widehat{\mu}^{(k)} = X^{(k)}\widehat{\beta}^{(k)}$, $\mu = (f(\frac{1}{n}), \ldots, f(1))^\top$. Zeige, dass $\widehat{\mu}^{(\widehat{k})}$ konsistent ist für μ, wenn m fest und $n \to \infty$.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 13.12.19.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

9. Übungsblatt

1. Betrachte für einen Klassifizierer $h : \mathcal{X} \to \{0, 1\}$ und $\alpha \in (0, 1)$ den Klassifikationsfehler

$$R_\alpha(h) = (1 - \alpha)\mathbb{P}(h(X) = 0, Y = 1) + \alpha\mathbb{P}(h(X) = 1, Y = 0).$$

   Bestimme unter Benutzung von $\eta(x) = \mathbb{P}(Y = 1|X = x)$ den Klassifizierer h^*_α der $R_\alpha(h^*_\alpha) = \min_h R_\alpha(h)$ erfüllt, wobei das Minimum über alle Klassifizierer genommen wird. Wie groß ist das Bayesrisiko $R^*_\alpha = R(h^*_\alpha)$?

2. Sei $h : \mathcal{X} \to \{0, 1\}$ ein Klassifizierer und $\eta(x) = \mathbb{P}(Y = 1|X = x)$. Zeige:

   (a) Es gilt

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{E}\big[\eta(X) + \mathbb{1}_{\{h(X)=1\}}(1 - 2\eta(X))\big].$$

   (b) Ist $h^*(x) = \mathbb{1}_{\{\eta(x)>1/2\}}$ der Bayes-Klassifizierer, so gilt

$$R(h) - R(h^*) = 2\mathbb{E}\big[|\eta(X) - 1/2|\mathbb{1}_{\{h(X)\neq h^*(X)\}}\big].$$

   (c) Für eine (feste) Funktion $\widehat{\eta} : \mathcal{X} \to [0, 1]$ betrachte nun den Plug-in-Klassifizierer $\widehat{h}(x) = \mathbb{1}_{\{\widehat{\eta}(x)>1/2\}}$. Dann gilt

$$R(\widehat{h}) - R(h^*) \leq 2\mathbb{E}\big[|\eta(X) - \widehat{\eta}(X)|\big].$$

   Interpretiere diese Schranke in Hinblick auf die Verbindung zwischen dem Klassifikationproblem und dem Problem der Regressionsschätzung.

3. Sei (X, Y) eine Paar von Zufallsvariablen mit Werten in $\mathbb{R}^p \times \{0, \dots, K - 1\}$. Zeige, dass die MAP-Regel $h^* : \mathbb{R}^p \to \{0, 1\}$ definiert durch

$$h^*(x) = \operatorname*{argmax}_{k=0,\dots,K-1} \mathbb{P}(Y = k|X = x)$$

   den Klassifikationsfehler $R(h) = \mathbb{P}(Y \neq h(X))$ über alle Klassifizierer $h : \mathbb{R}^p \to \{0, \dots, K - 1\}$ minimiert.

4. Betrachte Rosenblatts Perzeptron-Algorithmus:

   input: A training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, +1\}$
   initialize: $w^{(1)} = (0, \ldots, 0)$
     for $t = 1, 2, \ldots$
       if ($\exists i$ s.t. $y_i \langle w^{(t)}, x_i \rangle \leq 0$) then $w^{(t+1)} = w^{(t)} + y_i x_i$
         else output $w^{(t)}$

   Wir nehmen an, dass die Trainingsmenge linear separierbar ist, d.h. es gibt ein $w \in \mathbb{R}^p$ mit $y_i \langle w, x_i \rangle = y_i w^T x_i > 0$ für alle $i = 1, \ldots, n$. Des Weiteren nehmen wir an, dass $\|x_i\| \leq M$ für alle $i = 1, \ldots, n$. Sei $B = \min\{\|w\| : \forall i, y_i \langle w, x_i \rangle \geq 1\}$ und $w^* \in \mathbb{R}^p$ ein Vektor der das Minimum in der Definition von B annimmt. Zeige:

   (a) Läuft der Perzeptron-Algorithmus für T Schritte, so gilt

   $$\|w^{(T+1)}\|^2 \leq TM^2 \quad \text{und} \quad \langle w^*, w^{(T+1)} \rangle \geq T.$$

   (b) Schließen Sie mit Hilfe der Cauchy-Schwarz-Ungleichung, dass

   $$T \leq B^2 M^2.$$

   Der Perzeptron-Algorithmus stoppt also nach höchstens $B^2 M^2$ Schritten und liefert einen linearen Klassifizierer $h(x) = \text{sign}(\langle w, x \rangle)$ welcher alle Beobachtungen aus der Trainingsmenge richtig klassifiziert.

   (c) Konvergiert der Perzeptron-Algorithmus auch für andere Startwerte?

5. *Freiwillig:* Bearbeite die praktische Aufgabe 5.2.5 aus dem Buchprojekt *Methoden der Statistik und des maschinellen Lernens.*

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 17.1.20.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

10. Übungsblatt

1. Seien $\varepsilon_1, \dots, \varepsilon_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mathbb{P}(\varepsilon_i = +1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$. Außerdem seien $t_1, \dots, t_M \in \mathbb{R}^n$. Setze $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T \in \mathbb{R}^n$. Beweise schrittweise, dass

$$\mathbb{E} \max_{j \le M} \langle t_j, \varepsilon \rangle \le \max_{j \le M} \|t_j\| \sqrt{2 \log M}.$$

   (a) Für alle $\lambda > 0$ gilt $\mathbb{E} \max_{j \le M} \langle t_j, \varepsilon \rangle \le \frac{1}{\lambda} \log \Big(\sum_{j=1}^M \mathbb{E} e^{\lambda \langle t_j, \varepsilon \rangle} \Big)$.

   (b) Es gilt $\mathbb{E} e^{x \varepsilon_i} = (e^x + e^{-x})/2 \le e^{x^2/2}$ für alle $x \in \mathbb{R}$.

   (c) Wir erhalten $\mathbb{E} \max_{j \le M} \langle t_j, \varepsilon \rangle \le \frac{1}{\lambda} \Big(\log M + \lambda^2 \max_{j \le M} \|t_j\|^2 / 2 \Big)$ und die Behauptung folgt durch geeignete Wahl von λ.

2. Betrachte eine endliche Menge $\mathcal{H} = \{h_1, \dots, h_M\}$ von Klassifizierern und einen zugehörigen ERM-Klassifizierer $\widehat{h}_n$. Verwende den Symmetrisierungstrick und Aufgabe 1 um zu zeigen, dass

$$\mathbb{E} R(\widehat{h}_n) \le \min_{j \le M} R(h_j) + 2 \sqrt{\frac{2 \log M}{n}}.$$

   Zusatzaufgabe: Beweise eine analoge Schranke im Fall der affinen Klassifikation, mit $\log M$ ersetzt durch $(p+1)\log(n+1)$.

3. Für gegebene $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, +1\}$ und $\lambda > 0$, betrachte

$$\widehat{w} \in \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \max\big(1 - y_i \langle w, x_i \rangle, 0\big) + \lambda \|w\|^2.$$

   Zeige, dass $\widehat{w}$ von der Form $\widehat{w} = \sum_{i=1}^n \widehat{\alpha}_i y_i x_i$ ist, wobei

$$\begin{cases} \widehat{\alpha}_i = 0, & \text{falls } y_i \langle \widehat{w}, x_i \rangle > 1, \\ \widehat{\alpha}_i = 1/(2\lambda n), & \text{falls } y_i \langle \widehat{w}, x_i \rangle < 1, \\ \widehat{\alpha}_i \in [0, 1/(2\lambda n)], & \text{falls } y_i \langle \widehat{w}, x_i \rangle = 1. \end{cases}$$

   Punkte x_i mit $\widehat{\alpha}_i > 0$ heißen auch *Stützvektoren (support vectors)*.

   Hinweis: $f(z) := \max(z, 0)$ ist nicht differenzierbar in $z = 0$, aber für den Differenzenquotienten gilt $(f(h) - f(0))/h \in [0, 1]$ für alle $h \neq 0$. Alternativ dürfen auch die Rechenregeln des Subdifferentials verwendet werden, die im Anhang der Gliederung aufgeführt sind.

4. Sei (X, Y) ein Paar von Zufallsvariablen mit Werten in $\mathbb{R}^p \times \{0, 1\}$ und Verteilung

$$\mathbb{P}(Y = k) = \pi_k \quad \text{und} \quad \mathbb{P}(X \in dx | Y = k) = f_k(x)\, dx, \quad k \in \{0, 1\}, x \in \mathbb{R}^p,$$

wobei $\pi_0 + \pi_1 = 1$ und f_0, f_1 zwei Dichten in $\mathbb{R}^p$ sind. Zeige:

(a) Der Bayes-Klassifizierer h^* ist gegeben durch $h^*(x) = \mathbb{1}_{\{\pi_1 f_1(x) > \pi_0 f_0(x)\}}$ und das zugehörige Bayes-Risiko erfüllt

$$R^* = \int_{\mathbb{R}^p} \min(\pi_0 f_0(x), \pi_1 f_1(x))\, dx.$$

Seien f_0 und f_1 nun gegeben durch die Normalverteilungsdichte

$$f_k(x) = (2\pi)^{-p/2} \sqrt{\det(\Sigma_k^{-1})} \exp\Big(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\Big), \quad k = 0, 1,$$

mit invertierbaren Kovarianzmatrizen $\Sigma_k \in \mathbb{R}^{p \times p}$ und Mittelwerten $\mu_k \in \mathbb{R}^p$.

(b) Gilt $\Sigma_0 = \Sigma_1 = \Sigma$ und $\mu_0 \neq \mu_1$, so ist die Bedingung $\pi_1 f_1(x) > \pi_0 f_0(x)$ äquivalent zu

$$(\mu_1 - \mu_0)^T \Sigma^{-1} \Big(x - \frac{\mu_0 + \mu_1}{2} \Big) > \log(\pi_0 / \pi_1).$$

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 24.1.20.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

11. Übungsblatt

1. Sei $\mathcal{H} = \{h_1, \dots, h_M\}$ eine endliche Menge von Klassifizierern mit der Eigenschaft, dass $R(h_j) = 0$ für ein $j \leq M$. Des Weiteren sei $\widehat{h}_n$ ein zugehöriger ERM-Klassifizierer. Beweise schrittweise, dass

$$\mathbb{E}R(\widehat{h}_n) \leq \frac{1 + \log M}{n}.$$

   (a) Es gilt $R_n(\widehat{h}_n) = 0$ fast sicher.

   (b) Es folgt $\mathbb{P}(R(\widehat{h}_n) > \varepsilon) \leq \sum_{k:R(h_k)>\varepsilon} \mathbb{P}(R_n(h_k) = 0) \leq M(1-\varepsilon)^n$.

   (c) Die Behauptung folgt aus (b) und $\mathbb{E}R(\widehat{h}_n) = \int_0^\infty \mathbb{P}(R(\widehat{h}_n) > \varepsilon)\, d\varepsilon$.

   Man spricht von einer schnellen Rate (*fast rate*).

2. Eine lineare Abbildung $P : \mathbb{R}^p \to \mathbb{R}^p$ heißt Orthogonalprojektion falls (i) P ist idempotent, d.h. $P^2 = P$ und (ii) P ist symmetrisch, d.h. $P = P^T$.

   (a) Sei P eine Orthogonalprojektion und $\operatorname{Bild} P = \{Px : x \in \mathbb{R}^p\}$ das Bild von P. Zeige, dass $\langle x - Px, y\rangle = 0$ für alle $x \in \mathbb{R}^p$ und alle $y \in \operatorname{Bild} P$. Schließe, dass $\|x - Px\|^2 \leq \|x - y\|^2$ für alle $x \in \mathbb{R}^p$ und alle $y \in \operatorname{Bild} P$.

   (b) Sei V ein d-dimensionaler linearer Unterraum von $\mathbb{R}^p$. Dann gibt es eine eindeutig bestimmte Orthogonalprojektion P mit $\operatorname{Bild} P = V$. Ist $v_1, \dots, v_d$ eine Orthonormalbasis von V, so kann diese geschrieben werden als $P = \sum_{j=1}^d v_j v_j^T = (v_1 \cdots v_d)(v_1 \cdots v_d)^T$.

3. Für $X_1, \dots, X_n \in \mathbb{R}^p$ und $d \leq p$, betrachte das Optimierungsproblem

$$\min_{\mu,(z_i),V} \sum_{i=1}^n \|X_i - \mu - Vz_i\|^2 \tag{1}$$

   über $\mu \in \mathbb{R}^p$, $z_1, \dots, z_n \in \mathbb{R}^d$ mit $\sum_{i=1}^n z_i = 0$ und $V = (v_1 \cdots v_d) \in \mathbb{R}^{p\times d}$ mit $v_1, \dots, v_d \in \mathbb{R}^p$ Orthonormalsystem. Zeige:

   Eine Lösung $(\widehat{\mu}, (\widehat{z}_i), \widehat{V})$ ist von der Form $\widehat{\mu} = \bar{X} = n^{-1}\sum_{i=1}^n X_i$, $\widehat{z}_i = \widehat{V}^T(X_i - \bar{X})$ und $\widehat{V}$ ist eine Lösung des Optimierungsproblems

$$\min_V \sum_{i=1}^n \|X_i - \bar{X} - VV^T(X_i - \bar{X})\|^2 \tag{2}$$

   über $V = (v_1 \cdots v_d) \in \mathbb{R}^{p\times d}$ mit $v_1, \dots, v_d \in \mathbb{R}^p$ Orthonormalsystem. Wie sieht eine Lösung von (2) aus?

4. Seien $X_1, \ldots, X_n \in \mathcal{X}$ und $\Phi : \mathcal{X} \to \mathbb{R}^p$ eine Abbildung (*feature map*).

   (a) Wenden wir PCA auf $\Phi(X_1), \ldots, \Phi(X_n)$ an, so müssen wir zur Berechnung der ersten k Hauptkomponenten nur die k größten Eigenwerte und die dazugehörigen Eigenvektoren von

   $$K = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

   berechnen.

   (b) Wenden wir PCA auf die zentrierten Daten $\Phi(X_i) - n^{-1} \sum_{j=1}^n \Phi(X_j)$, $i = 1, \ldots, n$ an, so müssen wir zur Berechnung der ersten k Hauptkomponenten nur die k größten Eigenwerte und die dazugehörigen Eigenvektoren von

   $$K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n$$

   berechnen, wobei $(\mathbf{1}_n)_{ij} = 1/n$ für alle $i, j = 1, \ldots, n$.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 31.1.20.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

12. Übungsblatt

1. Sei S eine endliche Menge und $\mathcal{P}(S)$ die Potenzmenge von S. Zeige, dass die Abbildung $k : \mathcal{P}(S) \times \mathcal{P}(S) \to \mathbb{R}$ gegeben durch $k(A, B) = 2^{|A \cap B|}$ ein Kern ist. Hier bezeichnet $|A \cap B|$ die Anzahl der Elemente in dem Durchschnitt $A \cap B$.

2. Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein Kern. Zeige, dass die Abbildung $k' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ definiert durch $k'(x, y) = k(x, y)/\sqrt{k(x, x)k(y, y)}$, falls der Nenner ungleich Null ist und $k'(x, y) = 0$, sonst, auch eine Kern ist.

3. Seien V und W zwei d-dimensionale Unterräume von $\mathbb{R}^p$ mit Orthonormalbasen $v_1, \dots, v_d$ und $w_1, \dots, w_d$. Des Weiteren sei $A = (\langle v_j, w_k \rangle)_{j,k=1}^d$ die zugehörige Gramsche Matrix, und $A = \sum_{j=1}^d \sigma_j \psi_j \varphi_j^T$ eine Singulärwertzerlegung von A mit $1 \geq \sigma_1 \geq \cdots \geq \sigma_d \geq 0$ und Orthonormalbasen $\psi_1, \dots, \psi_d$ und $\varphi_1, \dots, \varphi_d$ von $\mathbb{R}^d$. Dann heißt

$$\vartheta_j := \vartheta_j(V, W) := \arccos(\sigma_j) \in [0, \pi/2]$$

der j-te Hauptwinkel (*principal angle*) zwischen V und W. Zeige:

   (a) Es gilt

$$\cos(\vartheta_1) = \max_{v \in V, w \in W} \frac{|\langle v, w \rangle|}{\|v\| \|w\|}.$$

   Formuliere und beweise eine analoge Formel für $\cos(\vartheta_j)$, $j = 2, \dots, d$.
   Hinweis: Verwende, dass $\sigma_1 = \max_{\|x\|, \|y\|=1} y^T A x$ und iterativ eine analoge Darstellung für die weiteren Singulärwerte.

   (b) Sind P_V und P_W die Orthogonalprojektionen auf V und W, so gilt

$$\|P_V - P_W\|_{\mathrm{HS}}^2 = 2 \sum_{j=1}^d \sin^2(\vartheta_j).$$

4. *Freiwillig:* Sind H_1 und H_2 zwei RKHS mit reproduzierenden Kernen $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, so ist $H_1 + H_2$ ein RKHS mit reproduzierendem Kern $k_1 + k_2$. Bestimme den RKHS von $k_1 + k_2$ in dem Fall, dass $\mathcal{X} = [0, 1]$, $k_1(x, y) = \min(x, y)$ und $k_2(x, y) = 1$.

Abgabe in Zweier- oder Dreier-Gruppen erfolgt vor der Vorlesung am Freitag, 7.2.20.

Vorlesung Methoden der Statistik
Wintersemester 2019/20

Randolf Altmeyer
Martin Wahl

Probeklausur zum 2. Teil der Vorlesung

1. Seien $(X_1, Y_1), \dots, (X_n, Y_n), (X, Y)$ i.i.d. Zufallsvariablen mit Werten in $\mathbb{R}^p \times \{\pm 1\}$. Es gelte $\|X\| \leq M$ fast sicher. Für $w \in \mathbb{R}^p$ mit $\|w\| \leq \lambda$ betrachte $R^{\text{hinge}}(w) = \mathbb{E}(1 - Y\langle w, X\rangle)_+$ und einen Minimierer $w^* \in \operatorname{argmin}_{\|w\|\leq\lambda} R^{\text{hinge}}(w)$, sowie $R_n^{\text{hinge}}(w) = \frac{1}{n}\sum_{i=1}^n (1 - Y_i\langle w, X_i\rangle)_+$ und einen Minimierer $\widehat{w}_n \in \operatorname{argmin}_{\|w\|\leq\lambda} R_n^{\text{hinge}}(w)$.

   (a) Erklären Sie im obigen Beispiel die Begriffe Trainingsmenge, ERM, konvexe Relaxation und Generalisierungsfehler.

   (b) Zeigen Sie, dass

$$\mathbb{E}R(\widehat{w}_n) - R(w^*) \leq \mathbb{E} \sup_{\|w\|\leq\lambda} (R(w) - R_n(w)).$$

   (c) Seien $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. Rademacher Zufallsvariablen, unabhängig von $(X_1, Y_1), \dots, (X_n, Y_n)$. Erklären Sie die folgenden Rechenschritte:

$$\begin{aligned}\mathbb{E} \sup_{\|w\|\leq\lambda} (R(w) - R_n(w)) &\overset{(1)}{\leq} 2\mathbb{E} \sup_{\|w\|\leq\lambda} \frac{1}{n}\sum_{i=1}^n \varepsilon_i (1 - Y_i\langle w, X_i\rangle)_+ \\ &\overset{(2)}{\leq} 2\mathbb{E} \sup_{\|w\|\leq\lambda} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle w, X_i\rangle \overset{(3)}{\leq} \frac{2\lambda M}{\sqrt{n}}.\end{aligned}$$

   Formulieren Sie eine Schranke für den Term $\mathbb{E}R(\widehat{w}_n) - R(w^*)$ aus (b).

2. Sei $f : \mathbb{R}^n \to [0, \infty)$ eine stetige Funktion und $\lambda > 0$. Für $x_1, \dots, x_n \in \mathbb{R}^p$ betrachte das Minimierungsproblem

$$\min_{w\in\mathbb{R}^p} \Big(f(\langle w, x_1\rangle, \dots, \langle w, x_n\rangle) + \lambda\|w\|^2 \Big). \tag{0.1}$$

   (a) Zeigen Sie, dass (0.1) eine Lösung besitzt.
   Hinweis. Man kann sich auf die Menge $\{w \in \mathbb{R}^p : \|w\| \leq \sqrt{f(0, \dots, 0)/\lambda}\}$ einschränken.

   (b) Ist w eine Lösung von (0.1), so existiert ein $\alpha \in \mathbb{R}^n$ mit $w = \sum_{j=1}^n \alpha_j x_j$.

   (c) Schreiben Sie (0.1) als ein Minimierungsproblem über $\alpha \in \mathbb{R}^n$ um, für das man nur die Einträge der Matrix $G = (\langle x_i, x_j\rangle)_{i,j=1}^n$ kennen muss.

   (d) Welche Wahl von f führt zum Ridge-Regression-Schätzer, welche zum SVM-Klassifizierer?

3. Sei (X, Y) ein Paar von Zufallsvariablen mit Werten in $\mathbb{R}^p \times \{0, 1\}$ und Verteilung gegeben durch

$$\mathbb{P}(Y = k) = \pi_k \quad \text{und} \quad \mathbb{P}(X \in dx | Y = k) = f_k(x)\, dx, \quad k \in \{0, 1\}, x \in \mathbb{R}^p,$$

mit $\pi_0 \in (0, 1)$, $\pi_0 + \pi_1 = 1$, und Normalverteilungsdichten

$$f_k(x) = (2\pi)^{-p/2} \sqrt{\det(\Sigma^{-1})} \exp\Big(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \Big),$$

wobei $\Sigma \in \mathbb{R}^{p \times p}$ symmetrisch und positiv definit und $\mu_0 \neq \mu_1 \in \mathbb{R}^p$.

(a) Zeigen Sie mit Hilfe der Bayesformel, dass

$$\mathbb{P}(Y = 1 | X = x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)} \qquad \mathbb{P}^X\text{-f.ü.}$$

(b) Schließen Sie, dass es $\alpha \in \mathbb{R}$ und $\beta \in \mathbb{R}^n$ gibt mit

$$\log\Big(\frac{\mathbb{P}(Y = 1 | X = x)}{\mathbb{P}(Y = 0 | X = x)} \Big) = \alpha + \beta^T x.$$

(c) Geben Sie den LDA-Klassifizierer und den Logistische-Regression-Klassifizierer an und diskutieren Sie kurz Unterschiede und Gemeinsamkeiten.

4. Für $X_1, \ldots, X_n \in \mathbb{R}^p$ und $d \leq p$ betrachte das Minimierungsproblem

$$\text{minimiere} \quad \sum_{i=1}^{n} \|X_i - UWX_i\|^2 \qquad \text{über } W \in \mathbb{R}^{d \times p}, U \in \mathbb{R}^{p \times d}.$$

(a) Interpretieren Sie W als Kompressionsmatrix und U als Wiederherstellungsmatrix.

(b) Zeigen Sie: das Minimierungsproblem besitzt eine Lösung von der Form $U = (u_1 \cdots u_d)$ mit $u_1, \ldots, u_d$ Orthonormalsystem in $\mathbb{R}^p$ und $W = U^T$.

(c) Definieren Sie die (unzentrierte) empirische Kovarianzmatrix $\widehat{\Sigma}$ und erklären Sie wie eine Lösung U aus (b) mit dieser zusammenhängt.