Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

1. und 2. Übungsblatt

1. Wir betrachten eine gleichmäßig auf dem Intervall $[a, b]$ verteilte mathematische Stichprobe $X_1, \ldots, X_n$ (d.h. $X_1, \ldots, X_n \sim U([a, b])$ i.i.d.) mit unbekannten Parametern $-\infty < a < b < \infty$.

   (a) Formalisieren Sie das statistische Modell.

   (b) Bestimmen Sie Maximum-Likelihood-Schätzer für a und b.

   (c) Welchen MSE haben die Maximum-Likelihood-Schätzer?

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 $N_1, \ldots, N_n$ der vergangenen n Tage.

   (a) Begründen Sie, weshalb $N_1, \ldots, N_n$ näherungsweise als unabhängig und Poiss(λ)-verteilt mit unbekanntem Parameter $\lambda > 0$ angesehen werden können. Geben Sie das entsprechende statistische Modell an.

   (b) Untersuchen Sie in dem Modell den Schätzer $\hat{\lambda} = (N_1 + \cdots + N_n)/n$ auf Erwartungstreue, Konsistenz und MSE. Weisen Sie nach, dass $\hat{\lambda}$ ML-Schätzer von λ ist.

   (c) Zeigen Sie für große n unter Verwendung der Normalapproximation und des $N(0, 1)$-Quantils q_α, dass

   $$I = \left[0, \hat{\lambda} + \frac{q_{0,95}}{\sqrt{n}} \sqrt{\hat{\lambda} + \frac{q_{0,95}^2}{4n}} + \frac{q_{0,95}^2}{2n}\right]$$

   approximativ ein einseitiges 95%-Konfidenzintervall für λ ist. Für $n \to \infty$ und

   $$\tilde{I}_n = \left[0, \hat{\lambda} + \frac{\sqrt{\hat{\lambda}}\, q_{0,95}}{\sqrt{n}}\right]$$

   gilt $\lim_{n \to \infty} \mathbb{P}_\lambda(\lambda \in \tilde{I}_n) = 0,95$, $\lambda > 0$ beliebig ($\tilde{I}_n$ ist asymptotisches 95%-Konfidenzintervall).
   Hinweis: Sie dürfen das Lemma von Slutsky aus Stochastik II bzw. der Vorlesung verwenden.

3. Die Beta-Verteilung $B(a,b)$ auf $[0,1]$ ist gegeben durch die Dichte

$$f_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in (0,1),$$

wobei $a,b > 0$ und Γ die Gamma-Funktion bezeichnet. $B(a,b)$ hat Erwartungswert $\mu_{a,b} = \frac{a}{a+b}$ und Varianz $\sigma^2_{a,b} = \frac{ab}{(a+b)^2(a+b+1)}$.

   (a) Skizzieren Sie $f_{a,b}$ für $(a,b) \in \{0.5; 1; 10\}^2$ (Computereinsatz gestattet).

   (b) Die Beobachtung X sei $\mathrm{Bin}(n,p)$-verteilt, wobei $n \geqslant 1$ bekannt ist sowie der unbekannte Parameter p gemäß $B(a,b)$ a priori verteilt ist. Zeigen Sie, dass der Bayes-optimale Schätzer (bzgl. MSE und $B(a,b)$) gegeben ist durch

$$\hat{p}_{a,b} = \frac{a+X}{a+b+n}.$$

   (c) Bestimmen Sie den MSE $R(p, \hat{p}_{a,b})$ als Funktion von p. Bestimmen Sie $a^*, b^* > 0$ so, dass $\max_{p\in[0,1]} R(p, \hat{p}_{a^*,b^*})$ minimal ist.

   (d) Zeigen Sie, dass $R(p, \hat{p}_{a^*,b^*})$ konstant in p ist. Schließen Sie, dass $\hat{p}_{a^*,b^*}$ minimax-Schätzer ist und der ML-Schätzer $\hat{p} = \frac{X}{n}$ nicht minimax ist.

4. Unter N Wahlberechtigten gibt es N_{AfD} AfD-Wähler sowie jeweils $N/2$ Männer bzw. Frauen.

   (a) Für gerades $n \in \mathbb{N}$ ziehe jeweils zufällig (mit Zurücklegen) $n/2$ Männer und $n/2$ Frauen und bestimme jeweils die Anzahl X_m bzw. X_w von AfD-Wählern/innen in den Stichproben. Formalisieren Sie das statistische Modell und zeigen Sie, dass $\hat{p}_{m/w} := \frac{X_m + X_w}{n}$ ein erwartungstreuer Schätzer von $p = N_{AfD}/N$ ist, dessen MSE nicht größer ist als der von $\hat{p} = \frac{X}{n}$ aus der Vorlesung.

   (b) Angenommen die Anzahl der AfD-Wähler unter den Männern ist etwa dreimal so groß wie der unter den Frauen. Wie können Sie durch geeignete Stichprobenzahlen einen erwartungstreuen Schätzer von p konstruieren, der noch kleineren MSE als $\hat{p}_{m/w}$ besitzt? (Stichwort: *stratified sampling*).

5. Jede Woche werden n Wahlberechtigte gefragt, ob Sie momentan CDU/CSU wählen würden. Die Anzahl der CDU/CSU-Wähler in Woche $i = 1, \dots, I$ sei als X_i notiert.

   (a) Unter welchen Modell-Voraussetzungen kann man $X_i \sim \mathrm{Bin}(n, p_i)$ mit $p_i \in [0,1]$ unbekannt sowie $(X_i)_{1 \leqslant i \leqslant I}$ unabhängig annehmen?

   (b) Es gelte das Modell in (a). Zeigen Sie, dass $\hat{p} = \frac{1}{I} \sum_{i=1}^{I} \frac{X_i}{n}$ erfüllt: $\mathbb{E}_{p_1,\dots,p_I}[\hat{p}] = p := \frac{1}{I} \sum_{i=1}^{I} p_i$, $\mathrm{Var}_{p_1,\dots,p_I}(\hat{p}) \leqslant \frac{p(1-p)}{nI}$.

   (c) Nun sei $I \in \mathbb{N}$ eine gerade Zahl sowie $\hat{v} = \frac{2}{I} \sum_{i=1}^{I/2} \frac{(X_{2i} - X_{2i-1})^2}{n^2}$ ein Schätzer der Fluktuationen von X_i/n. Weisen Sie nach, dass $\mathbb{E}_{p_1,\dots,p_I}[\hat{v}] \geqslant 2\frac{p(1-p)}{n}$ gilt. Erklären Sie damit in eigenen Worten die Umfragedifferenzwertanalyse unter `www.wahlrecht.de/lexikon/studie.html`.

6. Praktische Aufgabe.

   (a) Lesen Sie Kapitel 2.3 “Introduction to R” im Buch “Introduction to Statistical Learning” aus der Literaturliste, um sich mit den grundlegenden Funktionen von R vertraut zu machen. Die dort verwendete Datei “Auto.data” finden Sie unter `www-bcf.usc.edu/~gareth/ISL/Auto.data`.

   (b) Bearbeiten Sie Aufgabe 2.4.10 aus demselben Buch.

Aufgaben 1-3 werden korrigiert. Die Abgabe (in Zweier-Gruppen) erfolgt **vor** der Vorlesung am Freitag, 3.11.17.
Die Aufgaben 4-6 sind für die Übung am Dienstag, 7.11.17, schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

3. Übungsblatt

1. Beweisen Sie den *Korrespondenzsatz* für Tests und Konfidenzbereiche:

   (a) Ist $C_{1-\alpha}$ eine $(1-\alpha)$-Konfidenzmenge für $\vartheta \in \Theta$, so ist für jedes $\vartheta_0 \in \Theta$
   $$\varphi^{(\vartheta_0)}(x) = \mathbf{1}(\vartheta_0 \notin C_{1-\alpha}(x))$$
   ein Test vom Niveau α für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

   (b) Ist andererseits $(\varphi^{(\vartheta_0)})_{\vartheta_0 \in \Theta}$ eine Familie von nicht-randomisierten Tests für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$ zum Niveau α, so ist
   $$C_{1-\alpha}(x) = \{\vartheta_0 \in \Theta \,|\, \varphi^{(\vartheta_0)}(x) = 0\}$$
   eine $(1-\alpha)$-Konfidenzmenge.

   (c) Welche Form hat die Konfidenzmenge in (b), wenn die $\varphi^{(\vartheta_0)}$ jeweils Likelihood-Quotienten-Tests sind? Nehmen Sie der Einfachheit halber an, dass Θ ein metrischer Raum ohne isolierte Punkte und die Likelihoodfunktion L in ϑ stetig ist.

2. Es sei X $\mathrm{Bin}(n,p)$-verteilt mit $p \in [0,1]$ unbekannt. Zeigen Sie:

   (a) Für $p_0 \in (0,1)$ und das Testproblem $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ hat der Likelihood-Quotienten-Test die Form
   $$\varphi_\alpha = \mathbf{1}\Big(\frac{\hat{p}^X(1-\hat{p})^{n-X}}{p_0^X(1-p_0)^{n-X}} > c_\alpha\Big) + \gamma_\alpha \mathbf{1}\Big(\frac{\hat{p}^X(1-\hat{p})^{n-X}}{p_0^X(1-p_0)^{n-X}} = c_\alpha\Big)$$
   mit $\hat{p} = X/n$ und $c_\alpha \geqslant 0$, $\gamma_\alpha \in [0,1]$ geeignet.

   (b) Eine Münze wird zehnmal unabhängig geworfen und die Anzahl von 'Kopf' notiert. Bestimmen Sie einen Likelihood-Quotienten-Test für die Hypothese einer fairen Münze ($p_0 = 0,5$), der das Niveau $\alpha = 0,05$ exakt ausschöpft.

   (c) 2015 wurden in Berlin 38 030 Kinder geboren, darunter 19 614 Jungen. Die Wahrscheinlichkeit einer Jungengeburt sei p. Bestimmen Sie für den Test auf $H_0 : p = 0,5$ die kritischen Werte mittels Normalapproximation und geben Sie für die Daten von 2015 den p-Wert an.

(d) Bei einer Wahlbefragung unter 1 000 Wahlberechtigten erhalten die Grünen 6,4% der Stimmen. Welchen p-Wert ergibt ein Test auf die Null-Hypothese, dass die Grünen an der 5%-Hürde scheitern (mit Normalapproximation)?

3. Betrachten Sie das lineare Modell $Y = X\beta + \varepsilon$, wobei die Fehler ε $N(0, \Sigma)$-verteilt seien. Weisen Sie nach, dass der Kleinste-Quadrate-Schätzer $\hat{\beta}$ Maximum-Likelihood-Schätzer ist. Bestimmen Sie im Fall $\Sigma = \sigma^2 E_n$ mit $\sigma > 0$ unbekannt einen Maximum-Likelihood-Schätzer für (β, σ). Auf welche Schätzmethode führt der ML-Ansatz im Fall $\varepsilon_i \sim f_\varepsilon$ i.i.d. mit der Laplace-Dichte $f_\varepsilon(x) = \frac{1}{2\sigma} e^{-|x|/\sigma}$, $x \in \mathbb{R}$?

4. Praktische Aufgabe.

   (a) Lesen Sie Kapitel 3.6 "Lab: Linear Regression" im Buch "Introduction to Statistical Learning".

   (b) Bearbeiten Sie Aufgabe 3.7.15 aus demselben Buch.

Die Abgabe von Aufgaben 1-2 erfolgt **vor** der Vorlesung am Dienstag, 14.11.17.
Die Aufgaben 3-4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

4. Übungsblatt

1. Es sei $Y = X\beta + \varepsilon$ ein gewöhnliches lineares Modell mit $\varepsilon \sim N(0, \sigma^2 E_n)$. Wir untersuchen den Bayesansatz unter der a-priori-Verteilung $\beta \sim N(0, \tau^2 E_p)$, wobei $\sigma, \tau > 0$. Zeigen Sie:

   (a) Die a-posteriori-Verteilung des Parameters β ist $\beta \sim N(\mu_\beta, \Sigma_\beta)$ mit $\Sigma_\beta = (\sigma^{-2} X^\top X + \tau^{-2} E_p)^{-1}$, $\mu_\beta = \sigma^{-2}\Sigma_\beta X^\top Y$. Inwiefern ist das Verhalten für $\tau \to 0$ bzw. $\sigma \to 0$ jeweils plausibel?

   (b) Der Bayesschätzer ist ein *penalisierter* Kleinste-Quadrate-Schätzer (auch *ridge regression estimator* genannt)

   $$\hat{\beta}_{Bayes} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \Big(|Y - X\beta|^2 + \frac{\sigma^2}{\tau^2} |\beta|^2 \Big).$$

   (c) Ist λ_p der kleinste Eigenwert von $X^\top X$, so ist λ_p^{-1} der größte Eigenwert von der Kovarianzmatrix $\operatorname{Cov}(\hat{\beta})$ des Kleinste-Quadrate-Schätzers sowie $\lambda_p(\lambda_p + \sigma^2/\tau^2)^{-2}$ von $\operatorname{Cov}(\hat{\beta}_{Bayes})$. Was bedeutet das statistisch?

2. Betrachten Sie das statistische Modell

   $$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

   mit n verschiedenen Punkten $x_i \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ beliebig und unbekannt sowie $\mathbb{E}[\varepsilon] = 0$, $\operatorname{Cov}(\varepsilon) = \sigma^2 E_n$. Unter der polynomiellen Regressionsannahme $f = q_a^d$ mit

   $$q_a^d(x) = a_0 + a_1 x + \dots + a_d x^d$$

   sei $\hat{a} = (\hat{a}_0, \dots, \hat{a}_d)$ der Kleinste-Quadrate-Schätzer sowie $q_{\hat{a}}^d(x)$ die Vorhersage des Funktionswerts an der Stelle $x \in \mathbb{R}$. Zeigen Sie unter der Verwendung der *empirischen Norm* $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(x_i)^2$ einer Funktion $g : \mathbb{R} \to \mathbb{R}$:

   (a) Unter der Annahme $f = q_a^d$ gilt $\mathbb{E}_{q_a^d}[\|q_{\hat{a}}^d - q_a^d\|_n^2] = \frac{\sigma^2}{n}(d+1)$.
   Tipp: Betrachten Sie $|X\hat{\beta} - X\beta|^2$ im allgemeinen linearen Modell.

   (b) Für f beliebig im Modell ergibt sich die Bias-Varianz-Zerlegung

   $$\mathbb{E}_f[\|q_{\hat{a}}^d - f\|_n^2] = \min_{q_a^d} \|f - q_a^d\|_n^2 + \frac{\sigma^2}{n}(d+1),$$

   wobei sich der Bias durch den kleinsten Abstand von f zum Raum der Polynome vom maximalen Grad d in empirischer Norm ergibt.

(c) Geben Sie ein konkretes Beispiel an, wo f ein quadratisches Polynom ist, aber $q_{\hat{a}}^d$ für $d = 1$ kleineren Fehler im obigen Sinn besitzt als für $d = 2$.

3. Eine physikalische Größe $\mu \in \mathbb{R}$ wird in n verschiedenen Apparaturen gemessen. Für die Messwerte $Y_1, \ldots, Y_n$ nehmen wir $\mathbb{E}[Y_i] = \mu$, $\mathrm{Var}(Y_i) = \sigma_i^2$ mit $\sigma_i > 0$ bekannt und $\mathrm{Cov}(Y_i, Y_j) = 0$ für $i \neq j$ an. Schreiben Sie dies als lineares Modell und weisen Sie nach, dass das gewichtete Mittel

$$\hat{\mu} = \frac{\sum_{i=1}^n \sigma_i^{-2} Y_i}{\sum_{i=1}^n \sigma_i^{-2}}$$

Kleinste-Quadrate-Schätzer ist. Vergleichen Sie die Varianz von $\hat{\mu}$ mit der des Stichprobenmittels $\bar{Y}$ und beweisen Sie dann formal, dass $\hat{\mu}$ kleinste Varianz unter allen linearen erwartungstreuen Schätzern besitzt.

4. Melanom-Prognose: Lösen Sie Aufgabe 2.1.10.1 auf Seite 66 und Aufgabe 2.2.10.2 auf Seite 83 des Buchprojekts. Erklären Sie zuvor den Begriff des Prognoseintervalls.
Zusatzaufgabe: Lesen Sie Kapitel 2 bis einschließlich Abschnitt 2.2.2 und machen Sie Verbesserungsvorschläge.

Die Abgabe von Aufgaben 1-2 erfolgt **vor** der Vorlesung am Dienstag, 21.11.17.
Die Aufgaben 3-4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

5. Übungsblatt

1. Das Modell der einfaktoriellen Kovarianzanalyse (ANCOVA: analysis of covariance) mit $p \in \mathbb{N}$ Beobachtungsgruppen vom Umfang $n_1, \dots, n_p$ lautet

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

mit unbekannten Gruppenmittelwerten $\mu_1, \dots, \mu_p \in \mathbb{R}$, unbekanntem Regressionskoeffizienten $\kappa \in \mathbb{R}$, welcher die Abhängigkeit von einem Regressorvektor $x = (x_{ij}) \in \mathbb{R}^{\sum_{i=1}^p n_i}$ angibt und voneinander unabhängigen Störgrößen $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ mit (unbekannter) Varianz $\sigma^2 > 0$.

   (a) Bestimmen Sie die Designmatrix und den Parametervektor des zugehörigen linearen Modells und stellen Sie durch eine (notwendige und hinreichende) Bedingung an x sicher, dass die Designmatrix vollen Rang hat.

   (b) Bestimmen Sie den Kleinste-Quadrate-Schätzer $\hat{\beta} = (\hat{\mu}_1, \dots, \hat{\mu}_p, \hat{\kappa})^\top$.

   (c) Verifizieren Sie die Streuungszerlegung

$$\begin{aligned}\sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 =& \sum_{i=1}^p \sum_{j=1}^{n_i} \big(Y_{ij} - \hat{\kappa} x_{ij} - (Y_{i\bullet} - \hat{\kappa} x_{i\bullet})\big)^2 \\ &+ \sum_{i=1}^p n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \hat{\kappa}^2 \sum_{i=1}^p \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})^2\end{aligned}$$

   in eine Residuenvarianz innerhalb der Gruppen, eine Stichprobenvarianz zwischen den Gruppen, und eine Regressionsvarianz. Bestimmen Sie die Fisher-Statistik für einen Test der Hypothese $H_0 : \mu_1 = \dots = \mu_p$.

2. Es seien Y_i, $i = 1, \ldots, k$, unabhängige $\mathrm{Bin}(n_i, p_i)$-verteilte Beobachtungen mit $p_1, \ldots, p_k \in (0,1)$ unbekannt, $k \geqslant 2$. Betrachten Sie das Testproblem $H_0 : p_1 = \cdots = p_k$ gegen $H_1 : \exists j, l : p_j \neq p_l$. Beweisen Sie, dass für die Teststatistik $\mathcal{X}_n := \frac{4}{n} \sum_{i=1}^k (Y_i - \bar{Y})^2$ mit $n = \max(n_1, \ldots, n_k)$, $n_1(n), \ldots, n_k(n) \to \infty$ unter H_0 gilt
$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{X}_n > q_{1-\alpha}^{\chi^2(k-1)}\right) \leqslant \alpha, \quad \alpha \in (0,1).$$
Schließen Sie, dass der Test auf Homogenität in der Jungengeburtsrate aus der Vorlesung asymptotisch Niveau α besitzt.
Anleitung: Setzen Sie $Z_{i,n} = (Y_i - n_i(n)p_i)/(\sqrt{n_i(n)p_i(1-p_i)})$ und argumentieren Sie zunächst $\mathbb{P}(\sum_{i=1}^k (Z_{i,n} - \bar{Z}_{\bullet,n})^2 > q_{1-\alpha}^{\chi^2(k-1)}) \to \alpha$ mittels asymptotischer Normalität und Unabhängigkeit.
Freiwillig: Finden Sie die monatliche Geburtenstatistik für Deutschland (und ggf. weitere Länder) und überprüfen Sie, ob Sie die Annahme homogenen Jungenanteils zum Niveau 10% verwerfen können.

3. Untersuchen Sie im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ unter der Normalverteilungsannahme $\varepsilon \sim \mathcal{N}(0, \sigma^2 E_n)$ mit unbekanntem $\beta \in \mathbb{R}^p$ und $\sigma > 0$ folgende lineare Testprobleme:

   (a) Test auf signifikanten Einfluss einer Kovariablen auf die Responsevariable:
   $$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$
   für ein fest vorgegebenes $j \in \{1, \ldots, p\}$.

   (b) Test eines Subvektors $\beta^* = (\beta_1^*, \ldots, \beta_r^*)^\top \in \mathbb{R}^r$ mit $r \leqslant p$:
   $$H_0 : \beta_j = \beta_j^* \quad \forall j \in \{1, \ldots, r\} \quad \text{versus} \quad H_1 : \exists j \in \{1, \ldots, r\} : \beta_j \neq \beta_j^*.$$

   Stellen Sie diese Hypothesen in der Form $H_0 : K\beta = c$ mit geeigneter Kontrastmatrix K und einem Vektor c dar. Konstruieren Sie die zugehörigen Fisher-Statistiken und bestimmen Sie deren Verteilung.

4. Praktische Aufgabe: Klimaentwicklung.
Arbeiten Sie Beispiel 2.2.34 auf Seite 95 des Buchprojekts zu den jährlichen Temperaturdaten durch. Implementieren Sie die angesprochenen Testprobleme in einer Statistik-Software und bestimmen Sie die p-Werte. Verifizieren Sie die angegebenen Testentscheidungen.

Die Abgabe von Aufgaben 1-2 erfolgt **vor** der Vorlesung am Dienstag, 28.11.17.
Die Aufgaben 3-4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

6. Übungsblatt

1. Beweisen Sie den Satz aus der Vorlesung:
   Im verallgemeinerten linearen Modell mit kanonischer Linkfunktion und den Zeilen $x_1, \dots, x_n \in \mathbb{R}^p$ der Designmatrix gilt für die Loglikelihoodfunktion
   $$\nabla_\beta \ell(\beta, \varphi) = \frac{1}{\varphi} \sum_{i=1}^n (Y_i - \zeta'(\langle x_i, \beta \rangle)) x_i.$$
   Ist die Fisher-Informationsmatrix $I(\beta) = \frac{1}{\varphi} \sum_{i=1}^n \zeta''(\langle x_i, \beta \rangle) x_i x_i^\top$ positiv-definit für alle β und existiert ein $\hat{\beta}$ mit $\nabla_\beta \ell(\hat{\beta}, \varphi) = 0$ (für irgendein $\varphi > 0$), so ist $\hat{\beta}$ eindeutig bestimmter Maximum-Likelihood-Schätzer von β.

2. Weisen Sie nach, dass die Poissonverteilung mit Parameter $\lambda > 0$ eine Exponentialfamilie in $T(k) = k$ bildet mit natürlichem Parameter $\eta = \log \lambda \in \mathbb{R}$. Überprüfen Sie die Identitäten $\mathbb{E}_\eta[T] = \zeta'(\eta)$, $\mathrm{Var}_\eta(T) = \zeta''(\eta)$.

   Betrachten Sie nun das Modell der Poissonregression mit $\log \lambda_i = a x_i + b$ für gegebene Designpunkte $x_1, \dots, x_n \in \mathbb{R}$. Stellen Sie gemäß Aufgabe 1 eine Gleichung für den MLE auf und untersuchen Sie, ob der MLE existiert und eindeutig ist. Wie sieht die Iteration in Fishers Scoring-Methode aus?

3. Geben Sie die Cramér-Rao-Ungleichung samt Voraussetzungen aus der Literatur an. Weisen Sie damit für natürliche Exponentialfamilien nach, dass T erwartungstreuer Schätzer von minimaler Varianz für $\zeta'(\eta) = \mathbb{E}_\eta[T]$ für η im Innern des natürlichen Parameterbereichs ist.

4. *IRLS: Iteratively reweighted least squares.*
   Beweisen Sie, dass der Iterationsschritt in Fishers Scoring-Methode für logistische Regression mit Designmatrix X gegeben ist durch
   $$\hat{\beta}_{m+1} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} |W(\hat{\beta}_m)^{1/2} (Z(\hat{\beta}_m) - X\beta)|^2,$$
   wobei $Z(\beta) := X\beta + W(\beta)^{-1}(Y - p(\beta)) \in \mathbb{R}^n$, $p_i(\beta) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ mit $\eta = X\beta$ und $W(\beta) = \mathrm{diag}(p_i(\beta)(1 - p_i(\beta))_{1 \leqslant i \leqslant n})$.
   Wie kann man den Iterationsschritt jeweils als gewichtete Kleinste-Quadrate-Methode für ein lineares Modell mit Fehlerkovarianzmatrix Σ deuten?

Die Abgabe von Aufgaben 1-2 erfolgt **vor** der Vorlesung am Dienstag, 5.12.17.
Die Aufgaben 3-4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

7. Übungsblatt

1. Betrachten Sie für K Klassen mit Klassenwahrscheinlichkeiten $\pi_k \in [0,1]$, $\sum_{k=1}^K \pi_k = 1$, Mittelwerten $\mu_k \in \mathbb{R}^d$ und invertierbaren Kovarianzmatrizen $\Sigma_k \in \mathbb{R}^{d\times d}$ die Normalverteilungsmischungsdichte

$$f(x) = \sum_{k=1}^K \pi_k \varphi_{\mu_k,\Sigma_k}(x), \quad x \in \mathbb{R}^d.$$

Weisen Sie nach, dass die *decision boundary* zwischen Klasse k und ℓ (bei Standard-Klassifikationsfehler) gegeben ist durch $\{x \in \mathbb{R}^d \mid \delta_k(x) = \delta_\ell(x)\}$ mit quadratischen Diskrimanten

$$\delta_k(x) = -\tfrac{1}{2}\log(\det(\Sigma_k)) - \tfrac{1}{2}\langle \Sigma_k^{-1}(x-\mu_k), x-\mu_k\rangle + \log \pi_k, \quad k = 1,\dots,K.$$

Skizzieren Sie eine typische *decision boundary* im Fall $K = d = 2$.

2. Betrachten Sie das Optimierungsproblem

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^d} \Big(\tfrac{1}{2}|\beta|^2 + C\sum_{i=1}^n \big(1 - y_i(\langle x_i,\beta\rangle + \beta_0)\big)_+\Big)$$

für gegebene $y_i \in \{-1,+1\}$, $x_i \in \mathbb{R}^d$, $C > 0$. Beweisen Sie:
Die Lösung für β (deren Existenz vorausgesetzt sei) besitzt die Form

$$\beta = \sum_{i=1}^n \alpha_i y_i x_i \text{ mit } \alpha_i \in [0,C].$$

Punkte x_i mit $\alpha_i > 0$ heißen *Stützvektoren* (*support vectors*). Punkte x_i mit $y_i(\langle x_i,\beta\rangle + \beta_0) > 1$ sind keine Stützvektoren (sie liegen jenseits des *margin*). Für Punkte x_i mit $y_i(\langle x_i,\beta\rangle + \beta_0) < 1$ gilt $\alpha_i = C$ und diese x_i sind Stützvektoren.
Hinweis: $f(z) := z_+ = \max(z,0)$ ist nicht differenzierbar bei $z = 0$, aber für den Differenzenquotienten gilt $\frac{f(h)-f(0)}{h} \in [0,1]$ für alle $h \neq 0$.

3. Lesen Sie Abschnitt 4.4.5 *Logistic Regression or LDA?* im Buch *The Elements of Statistical Learning* (`https://web.stanford.edu/~hastie/Papers/ESLII.pdf`). Konstruieren Sie ein konkretes einfaches Beispiel (z.B. mit $K = 2$, $d = 1$), wo LDA besser klassifiziert als logistische Regression.

4. Praktische Aufgabe.

    (a) Lesen Sie Kapitel 4.6 “Lab: Logistic Regression, LDA, QDA and KNN” im Buch *Introduction to Statistical Learning*.

    (b) Bearbeiten Sie Aufgabe 4.7.10(a,d-h) aus demselben Buch.

    (c*) Probieren Sie, bessere Klassifikationsmethoden gemäß Aufgabe 4.7.10(i) zu finden.

Die Abgabe von Aufgaben 1-2 erfolgt **vor** der Vorlesung am Dienstag, 12.12.17. Aufgabe 3 ist für die Übung am 12.12., Aufgabe 4 für den 19.12. auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

8. Übungsblatt

1. Empirische Risikominimierung.
   Es seien $R, R_n : \mathscr{H} \to [0,\infty)$ Funktionen auf einer Menge $\mathscr{H}$ ($R(h)$ ist *Risiko* des statistischen Verfahrens h, $R_n(h)$ das empirische/geschätzte Risiko). Es gelte $\hat{h}_n \in \operatorname{argmin}_{h\in\mathscr{H}} R_n(h)$. Zeigen Sie für das Risiko des empirischen Risikominimierers $\hat{h}_n$ die Fundamentalungleichung:
   $$R(\hat{h}_n) \leqslant \inf_{h\in\mathscr{H}} R(h) + 2 \sup_{h\in\mathscr{H}} |R_n(h) - R(h)|.$$

2. Zur Beschreibung zeitlicher Entwicklungen mit deterministischer Wachstumstendenz und zufälligen Störungen verwendet man oft das folgende *autoregressive Modell* (der Ordnung 1):
   $$X_k = \gamma X_{k-1} + \varepsilon_k, \quad k = 1, \dots, n.$$
   Dabei sind $\gamma \in \mathbb{R}$ unbekannt, $X_0, \dots, X_n$ die Beobachtungen und $\varepsilon_1, \dots, \varepsilon_n$ unabhängige zufällige Störungen mit $\mathbb{E}[\varepsilon_k] = 0$ und $\operatorname{Var}(\varepsilon_k) = \sigma^2 > 0$.
   (a) Auf welchen Schätzer von γ führt ein Kleinste-Quadrate-Ansatz?
   (b) Unter welcher Verteilungsannahme an (ε_k) ist der Kleinste-Quadrate-Schätzer auch Maximum-Likelihood-Schätzer?
   (c) Ist dieser Schätzer erwartungstreu?

3. In einer Umfrage werden $n \in \mathbb{N}$ vierköpfige Familien nach der Anzahl ihrer Fußballfans befragt. Wir nehmen an, dass die Beobachtungen $Y_1, \dots, Y_n \in \{0, \dots, 4\}$ unabhängig voneinander $Bin(p_i, 4)$-verteilt sind mit unbekannten Parametern $p_1, \dots, p_n \in (0,1)$. Ein Fernsehsender identifiziert $k \in \mathbb{N}$ relevante Kovariablen und modelliert die Wahrscheinlichkeiten p_i über den Vektor $\eta = (\log \frac{p_1}{1-p_1}, \dots, \log \frac{p_n}{1-p_n})^\top \in \mathbb{R}^n$ mit $\eta = X\beta$ für eine Designmatrix $X \in \mathbb{R}^{n\times k}$ und einen unbekannten Parametervektor $\beta \in \mathbb{R}^k$. Weisen Sie nach, dass ein verallgemeinertes lineares Modell vorliegt. Was ist der natürliche Parameterraum? Wird die kanonische Linkfunktion verwendet?

Die Abgabe von Aufgabe 1 erfolgt **vor** der Vorlesung am Dienstag, 19.12.17.
Aufgaben 2-3 sind für die Übung schriftlich auszuarbeiten.

Wir wünschen Ihnen frohe Weihnachtstage und einen guten Start ins Jahr 2018!

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

9. Übungsblatt

1. Bestimmen Sie die Kullback-Leibler-Divergenz $\mathrm{KL}(N(\mu_1, \Sigma_1) \,|\, N(\mu_2, \Sigma_2))$ zwischen zwei d-dimensionalen Normalverteilungen mit $\mu_1, \mu_2 \in \mathbb{R}^d$ und symmetrischen, positiv-semidefiniten Matrizen $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$.

2. Wir betrachten ein Produktmodell $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$.

   (a) Zeigen Sie, dass die Kullback-Leibler-Divergenz für Produktmaße additiv ist:
   $$\mathrm{KL}(\mathbb{P}_\vartheta^{\otimes n} \,|\, \mathbb{P}_{\vartheta'}^{\otimes n}) = n\,\mathrm{KL}(\mathbb{P}_\vartheta \,|\, \mathbb{P}_{\vartheta'}) \quad \forall \vartheta, \vartheta' \in \Theta.$$

   (b) Es gelte außerdem $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta'}$ für alle $\vartheta, \vartheta' \in \Theta$ und $\ell(\vartheta, x) := \log L(\vartheta, x)$ sei die Loglikelihood-Funktion von $\mathbb{P}_\vartheta$ bezüglich einem $\mathbb{P}_{\vartheta^*}$. Weisen Sie für alle $\vartheta_0 \in \Theta$ nach, dass
   $$-\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, x_i) \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} \mathrm{KL}(\mathbb{P}_{\vartheta_0} \,|\, \mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P}_{\vartheta_0} \,|\, \mathbb{P}_{\vartheta^*}) \quad \text{für } n \to \infty$$
   gilt, sofern die Kullback-Leibler-Divergenzen endlich sind. Für welches $\vartheta \in \Theta$ ist die rechte Seite minimal?

   (c) Betrachten Sie die Konvergenz in (b) nun unter $\mathbb{P}^{\otimes n}$ für ein beliebiges Wahrscheinlichkeitsmaß $\mathbb{P}$ mit $\mathrm{KL}(\mathbb{P} \,|\, \mathbb{P}_\vartheta) < \infty$.
   Welcher Parameter in Θ ist also der natürliche Grenzwert des Maximum-Likelihood-Schätzers in einem *missspezifizierten Modell*, in dem die wahre Verteilung $\mathbb{P}$ nicht in $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ enthalten ist? (Nehmen Sie an, dass alles wohldefiniert ist und entsprechend konvergiert.)

3. Der *Totalvariationsabstand* zweier Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf einem messbaren Raum $(\mathcal{X}, \mathscr{F})$ ist definiert als

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

   (a) Weisen Sie $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int |p - q| \, \mathrm{d}\nu$, nach für ein dominierendes Maß ν und ν-Dichten $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$, $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$.

   (b) Zeigen Sie $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \leqslant \sqrt{\mathrm{KL}(\mathbb{P}|\mathbb{Q})/2}$.
   Hinweis: Betrachten Sie die Funktion $h(z) := z \log(z) - z + 1$, $z > 0$, mit stetiger Fortsetzung in Null. Zeigen Sie für alle $z \geqslant 0$:

$$\left(\tfrac{4}{3} + \tfrac{2}{3} z\right) h(z) \geqslant (z - 1)^2.$$

   Benutzen Sie anschließend die Cauchy-Schwarz-Ungleichung.

   (c) Prüfen Sie ob die Folgen $(\mathbb{P}_n)_{n \in \mathbb{N}}$ auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ schwach, im Totalvariationsabstand oder in der Kullback-Leibler-Divergenz gegen einen geeigneten Grenzwert $\mathbb{P}$ konvergieren (betrachten Sie sowohl $\mathrm{KL}(\mathbb{P}_n \,|\, \mathbb{P})$ als auch $\mathrm{KL}(\mathbb{P} \,|\, \mathbb{P}_n)$):

      (i) $\mathbb{P}_n = \delta_{1/n}$, Dirac-Maß in $1/n$;

      (ii) $\mathbb{P}_n = (1 - \frac{1}{n})\nu + \frac{1}{n}\delta_1$, wobei ν das Maß der $\mathcal{N}(0,1)$-Verteilung ist.

4. Praktische Aufgabe: AIC und BIC.

   (a) Vollziehen Sie die Beispiele 5.1.15 und 5.2.4 des Buchprojekts nach, um die Verwendung von AIC und BIC bei der Polynomregression zu verstehen.

   (b) Simulieren Sie ein Polynomregressionsmodell mit $x_i = i/n$, $i = 1, \ldots, n$, $\varepsilon_i \sim N(0,1)$ i.i.d., und Polynomfunktion $p(x) = x^3 - 2x^2 + x$. Bestimmen Sie für $n = 10$, $n = 100$ und $n = 1000$ in jeweils 100 Monte-Carlo-Iterationen:

      i. den *mittleren Vorhersage-Fehler* $E_k := \frac{1}{n} \sum_{i=1}^n (\hat{p}_k(x_i) - p(x_i))^2$ des geschätzten Polynoms $\hat{p}_k(x) = \sum_{j=0}^k \hat{\beta}_j x^j$ vom Grad k für $k = 0, 1, \ldots, 5$,

      ii. den von AIC und BIC gewählten Polynomgrad $\hat{k}^{AIC}$, $\hat{k}^{BIC}$ und die zugehörigen Fehler $E_{\hat{k}^{AIC}}$, $E_{\hat{k}^{BIC}}$ zusammen mit den Orakelfehlern $\min_k E_k$.

   Stellen Sie Ihre Ergebnisse mittels Boxplots über die Monte-Carlo-Iterationen dar. Wieso wird $\min_k E_k$ nicht stets bei $k = 3$ angenommen?

   (c) Freiwillig: Welche Ergebnisse in (b) ergeben sich, wenn die Daten mittels $Y_i = \sin(\pi x_i) + \varepsilon_i$ erzeugt werden?

Die Abgabe von Aufgaben 1 und 2 erfolgt **vor** der Vorlesung am Dienstag, 16.1.18. Aufgaben 3 und 4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

10. Übungsblatt

1. Betrachten Sie die Beobachtungen $Y = \mu + \varepsilon$ mit $\mu \in \mathbb{R}^n$ und $\varepsilon \sim N(0, \Sigma)$, wobei $\Sigma \in \mathbb{R}^{n \times n}$ bekannt und strikt positiv-definit sei. Leiten Sie (in Analogie zum Fall $\Sigma = \sigma^2 E_n$) für die Modelle $Y = X^{(k)}\beta^{(k)} + \varepsilon$ mit $\beta^{(k)} \in \mathbb{R}^k$ und $X^{(k)} \in \mathbb{R}^{n \times k}$ von vollem Rang k, $k = 1, \dots, K \leqslant n$, das Akaike-Informationskriterium AIC(k) her. Weisen Sie $\mathbb{E}[AIC(k)] = \mathbb{E}[d(\hat{\beta}_k)]$ nach.

2. Es X_p eine $\chi^2(p)$-verteilte Zufallsvariable mit $p \in \mathbb{N}$ Freiheitsgraden. Ziel ist eine Abschätzung von

$$\mathbb{P}\left(X_p - \mathbb{E}[X_p] \geqslant \kappa \sqrt{\mathrm{Var}(X_p)}\right), \quad \kappa > 0.$$

   (a) Berechen Sie $\mathbb{E}[X_p]$, $\mathrm{Var}(X_p)$, $\mathbb{E}[e^{\lambda X_p}]$ für $\lambda > 0$.

   (b) Schätzen Sie die Wahrscheinlichkeit mit der (verallgemeinerten) Markovungleichung von oben ab.

   (c) Minimieren Sie die Schranke durch eine optimale Wahl von λ.

3. Eine zufällige Designmatrix $X^{(n)} \in \mathbb{R}^{n \times k}$, $k \leqslant n$, sei gegeben durch i.i.d. Beobachtungspunkte $X_1, \dots, X_n \in D$, Ansatzfunktion $\varphi_1, \dots \varphi_k : D \to \mathbb{R}$ sowie $X_{ij}^{(n)} = \varphi_j(X_i)$. Wir beobachten $Y = X^{(n)}\beta + \varepsilon$ mit $\beta \in \mathbb{R}^k$, $\varepsilon \sim N(0, \sigma^2 E_n)$ (unabhängig vom Design). Zeigen Sie:

   (a) Gilt $\varphi_1(X_1), \dots, \varphi_k(X_1) \in L^2$, so konvergiert $\hat{\Sigma}_n := \frac{1}{n}(X^{(n)})^\top X^{(n)}$ für $n \to \infty$ fast sicher gegen $\Sigma := (\mathbb{E}[\varphi_i(X_1)\varphi_j(X_1)])_{1 \leqslant i,j \leqslant k}$. Im Fall $\Sigma > 0$ folgt insbesondere $\mathbb{P}(\mathrm{rank}(X^{(n)}) = k) \to 1$.

   (b) Mit der empirischen Norm $\|v\|_n^2 := \frac{1}{n}\sum_{i=1}^n v_i^2$, $v \in \mathbb{R}^n$, gilt $\mathbb{E}[\|X^{(n)}\beta\|_n^2] = \mathbb{E}[\sum_{j=1}^k \beta_j^2 \varphi_j(X_1)^2]$, $\beta \in \mathbb{R}^k$, unabhängig von n.

   (c) Für den KQ-Schätzer $\hat{\beta}_n$, so er existiert, gilt $\mathbb{E}[\|X^{(n)}(\hat{\beta}_n - \beta)\|_n^2] = \frac{\sigma^2 k}{n}$.

   (d) Es gilt $\mathbb{E}^\varepsilon[|\hat{\beta}_n - \beta|^2] = \frac{\sigma^2}{n}\,\mathrm{trace}(\hat{\Sigma}_n^{-1})$ (Erwartungswert nur bzgl. (ε_i)). Schließen Sie $\lim_{n \to \infty} n\,\mathbb{E}^\varepsilon[|\hat{\beta}_n - \beta|^2] = \sigma^2\,\mathrm{trace}(\Sigma^{-1})$ f.s. für $\Sigma > 0$.

   (e*) (freiwillig) Gilt auch $\lim_{n \to \infty} n\,\mathbb{E}[|\hat{\beta}_n - \beta|^2] = \sigma^2\,\mathrm{trace}(\Sigma^{-1})$ für $\Sigma > 0$?

4. Praktische Aufgabe: Random design regression.

   (a) Im Rahmen von Aufgabe 3 betrachten Sie $X_1, \dots, X_n \sim N(0,1)$ sowie $\varphi_j(x) = x^j$, $j = 0, 1, \dots, k$. Im wahren Modell gelte $Y_i = X_i^4 + \varepsilon_i$, $i = 1, \dots, n$. Bestimmen Sie für $n \in \{10, 100, 1000\}$, $k = 6$ in jeweils 100 Monte-Carlo-Iterationen den Vorhersagefehler $\|X^{(n)}(\hat{\beta}_n - \beta)\|_n^2$ und den Schätzfehler $|\hat{\beta}_n - \beta|^2$. Stellen Sie die Ergebnisse im Boxplot dar und vergleichen Sie mit den theoretischen Ergebnissen aus Aufgabe 3.

   (b) Führen Sie dieselben Rechnungen durch bei deterministischem Design $x_i = \Phi^{-1}(i/n)$, $i = 1, \dots, n$, wobei Φ^{-1} die Inverse der Verteilungsfunktion (d.h. die Quantilsfunktion) von $N(0,1)$ bezeichnet.

Die Abgabe von Aufgaben 1 und 2 erfolgt **vor** der Vorlesung am Dienstag, 23.1.18. Aufgaben 3 und 4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

11. Übungsblatt

1. (a) Beweisen Sie $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \tau)\,\mathrm{d}\tau$ für eine beliebige Zufallsvariable X mit Werten in $[0, \infty)$.

   (b) Folgern Sie aus dem Hauptsatz zur Modellwahl und unter dessen Voraussetzungen die Orakelungleichung

$$\mathbb{E}[\|\hat{\mu}_{\hat{m}} - \mu\|^2] \leqslant C(K, \kappa)\Big(\inf_m \big\{\|\mu_m - \mu\|^2 + \mathrm{pen}(d_m)\big\} + \sigma^2 \sum_{m=1}^M e^{-d_m\kappa^2/2}\Big),$$

   für eine Konstante $C(K, \kappa) > 0$.

2. In der multiplen linearen Regression

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i \quad \text{für} \quad i = 1, \ldots, n$$

   mit Kovariablen $(x_{1,\ell}, \ldots x_{n,\ell}), 1 \leqslant \ell \leqslant k$, unbekanntem Parametervektor $\beta \in \mathbb{R}^{k+1}$ und Fehlern $\varepsilon_1, \ldots, \varepsilon_n \overset{iid.}{\sim} \mathcal{N}(0, \sigma^2), \sigma^2 > 0$, soll eine Variablenselektion durchgeführt werden, um das Modell auf die aktiven Kovariablen zu reduzieren. Welches $\kappa > 0$ muss gewählt werden, um den Term

$$T := \sum_{m=1}^M e^{-d_m\kappa^2/2}$$

   aus Aufgabe 1 gleichmäßig zu beschränken? Gehen Sie wie folgt vor:

   (a) Zeigen Sie $M = 2^{k+1}$ und $T = \sum_{\ell=1}^{k+1} \binom{k+1}{\ell} e^{-\ell\kappa^2/2}$.

   (b) Schätzen Sie $\log \binom{k+1}{\ell}$ linear in ℓ ab und bestimmen Sie κ in Abhängigkeit von k.

3. Das *empirische Skalarprodukt* sei definiert als $\langle f,g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(\frac{i}{n})g(\frac{i}{n})$ bzw. $\langle x,g\rangle_n := \frac{1}{n}\sum_{i=1}^n x_i g(\frac{i}{n})$ für Funktionen f,g auf $[0,1]$ bzw. $x \in \mathbb{R}^n$. Setze $\|f\|_n^2 := \langle f,f\rangle_n$. Sei $(\varphi_k)_{k=1}^n$ eine Orthonormalbasis bezüglich $\langle \bullet, \bullet\rangle_n$.

   Für $f(x) = \sum_{k=1}^n \beta_k \varphi_k(x), x \in [0,1]$, mit Koeffizienten $\beta_1,\dots,\beta_n \in \mathbb{R}$ betrachten wir das Regressionsmodell

   $$Y_i = f\big(\tfrac{i}{n}\big) + \varepsilon_i, \quad i = 1,\dots,n,$$

   mit unabängig und identisch $\mathcal{N}(0,\sigma^2)$-verteilten Fehlern $\varepsilon_i, i = 1,\dots,n$.

   (a) Weisen Sie nach, dass $\hat\beta_k := \langle Y, \varphi_k\rangle_n$ Maximum-Likelihood-Schätzer von β_k für $k = 1,\dots,n$ ist.

   (b) Bestimmen Sie für $m \in \{1,\dots,n\}$ den Fehler

   $$\mathbb{E}\left[\|f - \hat f_m\|_n^2\right] \quad \text{für} \quad \hat f_m(x) := \textstyle\sum_{k=1}^m \hat\beta_k \varphi_k(x).$$

   Auf welches Minimierungsproblem führt eine optimale Wahl von m?

   (c) Nehmen Sie an, dass es ein $s > 0$ und $0 < c < C$ gibt, so dass $ck^{-s} \leqslant |\beta_k| \leqslant Ck^{-s}$ gilt. Wie wächst das optimale m in n?

4. Es sei ein zweidimensionaler Zufallsvektor (X,Y) mit Lebesguedichte $f^{(X,Y)}$ gegeben.

   (a) Bestimmen Sie die bedingte Wahrscheinlichkeit $P(X \leqslant x \mid Y \in [y-h, y+h])$ für $x,y \in \mathbb{R}, h > 0$ mit $f^Y(y) > 0$ und f^Y stetig bei y. Zeigen Sie, dass der Grenzwert der bedingten Wahrscheinlichkeit für $h \to 0$ existiert und gegeben ist durch

   $$\frac{\int_{-\infty}^x f^{(X,Y)}(\xi,y)\,d\xi}{f^Y(y)} =: F^{X|Y=y}(x).$$

   (b) Man definiert daher die *bedingte Dichte* von X gegeben $Y = y$ als

   $$f^{X|Y=y}(x) := \frac{\partial}{\partial x}F^{X|Y=y}(x) = \frac{f^{(X,Y)}(x,y)}{f^Y(y)}, \text{ sofern } f^Y(y) > 0.$$

   Zeigen Sie, dass $f^{X|Y=y}$ in der Tat eine Dichtefunktion ist für alle y mit $f^Y(y) > 0$ und die Bayesformel für Dichten gilt:

   $$f^{X|Y=y}(x) = \frac{f^{(X,Y)}(x,y)}{\int_{\{\xi : f^X(\xi)>0\}} f^{Y|X=\xi}(y) f^X(\xi)\,d\xi} \text{ sofern } f^Y(y) > 0.$$

   (c) Verallgemeinern Sie (b) auf Dichten $f^{(X,Y)}$ bezüglich einem allgemeinen Produktmaß $\mu_X \otimes \mu_Y$. Zeigen Sie ferner für $X \in L^1$, dass $\int x f^{X|Y=y}(x)\,\mu_X(dx)$ die Eigenschaften des bedingten Erwartungswerts $\mathbb{E}[X \mid Y = y]$ erfüllt.

Die Abgabe von Aufgaben 1 und 2 erfolgt **vor** der Vorlesung am Dienstag, 30.1.18. Aufgaben 3 und 4 sind für die Übung schriftlich auszuarbeiten.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

12. Übungsblatt: Probeklausur

1. Entscheiden Sie, ob folgende Aussagen wahr oder falsch sind.

   (a) Minimax-optimale Schätzer haben für keinen Parameterwert einen größeren Fehler als Bayesschätzer.

   (b) Die kritischen Werte im Gaußtest leiten sich aus der t-Verteilung her.

   (c) In der Varianzanalyse wird die Streuung zwischen den Gruppen und die innerhalb der Gruppen mit einem F-Test verglichen.

   (d) Sind im linearen Modell die Fehler normalverteilt, so ist die Stichprobenvarianz $\hat{\sigma}^2 = |Y - X\hat{\beta}_{\mathrm{KQ}}|^2/(n-k)$ χ^2-verteilt.

   (e) Die Verteilungen $(\mathrm{Bin}(n,p))_{p\in(0,1)}$ bilden eine Exponentialfamilie.

   (f) Für großen Stichprobenumfang n wählt AIC ein kleineres Modell als BIC.

   (g) Bei der logistischen Regression modelliert man $Y_i \sim \mathrm{Bin}(1, p_i)$ mit $p_i = \log((X\beta)_i)$, $i = 1, \dots n$.

   (h) Der Median von $(X_1, \dots, X_n)$ ist Maximum-Likelihood-Schätzer, wenn die X_i unabhängig sind mit Dichte $f(x) = \frac{\lambda}{2} e^{-\lambda|x|}$, $x \in \mathbb{R}$, wobei $\lambda > 0$ beliebig.

   (i) Der Unterschied zwischen den Klassifikationen mit *separating hyperplanes* und *support vectors* ist, dass im zweiten Fall falsch klassifizierte Trainingsdaten erlaubt sind, wenn dies einen signifikant größeren *margin* erlaubt.

   (j) Für Folgen $(P_n)_{n\ge1}$ und $(Q)_{n\ge1}$ von Maßen auf einem gemeinsamen messbaren Raum gilt: $\mathrm{TV}(P_n, Q_n) \to 0 \Rightarrow \mathrm{KL}(P_n|Q_n) \to 0$.

2. Es seien (X, Y) ein Zufallsvektor auf $\mathbb{R}^d \times \{0,1\}$ sowie $(X_1, Y_1,), \dots (X_n, Y_n)$ unabhängige, wie (X, Y) verteilte Beobachtungen. Für einen Klassifizierer $C : \mathbb{R}^d \to \{0,1\}$ betrachte den Klassifizierungsfehler $R(C) := \mathbb{P}(C(X) \neq Y)$.

   (a) Zeigen Sie, dass der Bayes-Klassifizierer $C_B(x) = \mathrm{argmax}_{k=0,1}\, \mathbb{P}(Y = k \,|\, X = x)$ den Klassifizierungsfehler minimiert. Wieso ist der Bayes-Klassifizierer für den Statistiker unzugänglich?

   (b) Auf welche Klassifikationsmethode führt der Ansatz, dass X gegeben $\{Y = k\}$ $N(\mu_k, \Sigma)$-verteilt mit $\mu_0 \neq \mu_1$ und $\Sigma > 0$ ist? Wie sieht die zugehörige Klassifikationsgrenze (*classification boundary*) im $\mathbb{R}^d$ geometrisch aus?

(c) Setze $d = 1$ und modelliere $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 1) = \pi_0 \in (0,1)$ sowie $f^{Y|X=k}(x) = \lambda_k e^{-\lambda_k x}$, $x \geqslant 0$, $k = 0,1$ (exponential-verteilt mit $0 < \lambda_0 < \lambda_1$). Benutzen Sie die Bayesformel, um den Bayesklassifizierer zu bestimmen. Schätzen Sie diesen dann mittels eines Maximum-Likelihood-Schätzers von $(\pi_0, \lambda_0, \lambda_1)$, basierend auf den Daten $(X_1, Y_1,), \ldots (X_n, Y_n)$.

3. Betrachten Sie das lineare Modell $Y = X\beta + \varepsilon$ mit $\beta \in \mathbb{R}^k$, $X \in \mathbb{R}^{n\times k}$ mit $\operatorname{rank}(X) = k \leqslant n$ und $\mathbb{E}[\varepsilon] = 0$, $\operatorname{Cov}(\varepsilon) = \sigma^2 E_n$. Definiere den linearen Schätzer
$$\hat\beta = AY \text{ mit } A \in \mathbb{R}^{k\times n} \text{ deterministisch.}$$

   (a) Bestimmen Sie Bias, Varianz und mittleren quadratischen Fehler (MSE) von $\hat\beta$ in Abhängigkeit von A.

   (b) Zeigen Sie für Vorhersagefehler und *residual sum of squares* $RSS = |X\hat\beta - Y|^2$ den allgemeinen Zusammenhang
$$\mathbb{E}[|X(\hat\beta - \beta)|^2] = \mathbb{E}[RSS] + 2\sigma^2 \operatorname{trace}(XA) - \sigma^2 n.$$

   (c) Leiten Sie auf diesem Weg für den Kleinste-Quadrate Schätzer ab, dass das AIC-Kriterium für die Modellwahl die Kullback-Leibler-Devianz erwartungstreu schätzt.

4. Betrachten Sie in Aufgabe 3 nun die *ridge-regression*-Schätzer $\hat\beta_\lambda = A_\lambda Y$ mit $A_\lambda = (X^\top X + \lambda E_n)^{-1} X^\top$, $\lambda > 0$.

   (a) Begründen Sie $\hat\beta_\lambda = \operatorname{argmin}_{b\in\mathbb{R}^k} |Y - Xb|^2 + \lambda |b|^2$. Ist $\hat\beta_\lambda$ auch im Fall $\operatorname{rank}(X) < k$ wohldefiniert?

   (b) Setze $v(\lambda) = \operatorname{Var}(\hat\beta_\lambda)$, $\lambda > 0$. Zeigen Sie, dass $v(\lambda)$ monoton fallend in λ ist mit $\lim_{\lambda\downarrow 0} v(\lambda) = \operatorname{Var}(\hat\beta)$ für den Kleinste-Quadrate-Schätzer $\hat\beta$ und $\lim_{\lambda\uparrow\infty} v(\lambda) = 0$. Wieso widerspricht das nicht dem Satz von Gauß-Markov?

   (c) Es seien $s_1 \geqslant s_2 \geqslant \cdots s_n \geqslant 0$ die Eigenwerte von $X^\top X$. Weisen Sie mit Aufgabe 3 nach, dass
$$R(\lambda) := |X\hat\beta_\lambda - Y|^2 + 2\sigma^2 \sum_{i=1}^n \frac{s_i}{s_i + \lambda} - n$$
ein unverzerrter Schätzer des Vorhersagefehlers $|X(\hat\beta_\lambda - \beta)|^2$ ist. Wie könnte man also den *Regularisierungsparameter* λ wählen?

Die Abgabe von Aufgaben 1-3 erfolgt **vor** der Vorlesung am Dienstag, 6.2.18.