Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

9. Übungsblatt

1. Beweisen Sie die *Höffding-Ungleichung*: Für positive Zahlen R_i sowie unabhängige und zentrierte Zufallvariablen X_i mit $|X_i| \leqslant R_i$ f.s. gilt

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^n R_i^2}\Big), \quad t \geqslant 0.$$

   Zeigen Sie dazu $e^{\alpha x} \leqslant \frac{R-x}{2R}e^{-\alpha R} + \frac{R+x}{2R}e^{\alpha R}$ für $\alpha \geqslant 0$ und $|x| \leqslant R$ und schließen Sie $\mathbb{E}[e^{\alpha X_i}] \leqslant e^{\alpha^2 R_i^2/2}$. Verwenden Sie die Markovungleichung in geeigneter Weise, um zunächst $\mathbb{P}(\sum_{i=1}^n X_i \geqslant t)$ abzuschätzen.

   Nun sei $X_i = Y_i - \mathbb{E}[Y_i]$ mit $Y_i \sim \mathrm{Bern}(p)$ für $p \in (0,1)$. Vergleichen Sie in diesem Fall die Schranke der Höffding-Ungleichung mit der der Tschebyschew-Ungleichung.

2. Für einen Klassifizierer C und eine mathematische Stichprobe (*Trainingsdaten*) $(X_i, Y_i)_{1\leqslant i\leqslant n}$ bezeichnet

$$R_n(C) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i \neq C(X_i))$$

   das sogenannte *empirische Risiko*. $\hat{C}$ heißt *ERM-Klassifizierer* (*empirical risk minimizer*) in einer Klasse $\mathscr{C}$ von Klassifizierern, falls $R_n(\hat{C}) = \min_{C\in\mathscr{C}} R_n(C)$ gilt. Zeigen Sie für das *Risiko* (den Klassifizierungsfehler) R die Fundamentalungleichung

$$R(\hat{C}) \leqslant \inf_{C\in\mathscr{C}} R(C) + 2\sup_{C\in\mathscr{C}} |R_n(C) - R(C)|.$$

3. Betrachten Sie eine endliche Familie $\mathscr{C} = \{C_1, \dots, C_M\}$ von Klassifizierern und den zugehörigen ERM-Klassifizierer $\hat{C}$. Verwenden Sie obige Ergebnisse, um für alle $\tau > 0$ zu zeigen, dass mit Wahrscheinlichkeit mindestens $1 - e^{-\tau}$

$$R(\hat{C}) \leqslant \min_{1\leqslant m\leqslant M} R(C_m) + \frac{\sqrt{8(\log(2M) + \tau)}}{\sqrt{n}}.$$

4. Wir betrachten die Klassifikation bei $K \in \mathbb{N}$ Klassen, es gilt also $Y \in \{1, \dots, K\}$. Zeigen Sie, dass bezüglich dem Klassifikationsfehler $R(C) = \mathbb{P}(C(X) \neq Y)$ eines Klassifizierers $C : \mathcal{X} \to \{1, \dots, K\}$ der Bayesklassifizierer

$$C^*(x) := \operatorname{argmax}_{k=1,\dots,K} \eta_k(x) \text{ mit } \eta_k(x) := \mathbb{P}(Y = k \mid X = x)$$

minimalen Fehler besitzt. Wie groß ist das entsprechende Bayes-Risiko?

Abgabe am Dienstag, 10.1.23, über *Moodle*.

Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

10. Übungsblatt

1. Weisen Sie nach, dass das Bayesrisiko im LDA-Modell gegeben ist durch

$$R^* = \pi_{-1}\Phi\Big(\big(\log(\tfrac{\pi_{+1}}{\pi_{-1}}) - \Delta^2/2\big)/\Delta\Big) + \pi_{+1}\Big(1 - \Phi\Big(\big(\log(\tfrac{\pi_{+1}}{\pi_{-1}}) + \Delta^2/2\big)/\Delta\Big)\Big)$$

   mit der $N(0,1)$-Verteilungsfunktion Φ und dem *Mahalanobisabstand* Δ. Diskutieren Sie das Verhalten jeweils für $\pi_{+1} \to 1$, $\Delta \to \infty$ und $\Delta \to 0$.

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

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

   wobei $\varphi_{\mu,\Sigma}$ die $N(\mu,\Sigma)$-Dichte bezeichne. Weisen Sie nach, dass der beste Klassifizierer (bei Standard-Klassifikationsfehler) gegeben ist durch $C^*(x) = \operatorname{argmax}_{k=1,\dots,K} \delta_k(x)$ mit quadratischen Diskrimanten

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

   Welche geometrischen Formen für $p = 2$ können die Entscheidungsgrenzen $\{x \in \mathbb{R}^2 \mid \delta_k(x) = \delta_l(x)\}$ zwischen Klasse k und l besitzen?

3. Beweisen Sie für den $\mathcal{O}_P$-Kalkül von reellwertigen Zufallsvariablen X_n, Y_n, X und Zahlen $a_n, b_n > 0$:

   (a) Aus $a_n^{-1} X_n \xrightarrow{d} X$ folgt $X_n = \mathcal{O}_P(a_n)$ (Konvergenz in Verteilung impliziert stochastische Beschränktheit).

   (b) Aus $X_n = \mathcal{O}_P(a_n)$, $Y_n = \mathcal{O}_P(b_n)$ folgt $X_n + Y_n = \mathcal{O}_P(a_n + b_n)$ und $X_n Y_n = \mathcal{O}_P(a_n b_n)$.

4. Klassifikation in der Praxis: Laden Sie die Cleveland-Daten (`https://archive.ics.uci.edu/ml/datasets/heart+disease`, vgl. Beispiel 5.1 im Buch) herunter. Benutzen Sie die ersten $n = 151$ Patienten mit den Kovariablen Ruheblutdruck und maximale Herzfrequenz und ihren Labels (0=gesund, 1-4=krank) als Trainingsdaten. Die restlichen $m = 150$ Patienten dienen als *Testdaten*.

   (a) Führen Sie eine logistische Regression für die Trainingsdaten durch (mit $p = 3$, d.h. unter Verwendung eines konstanten Glieds), zeichnen Sie die erhaltene Entscheidungsgrenze in das Koordinatensystem und bestimmen Sie den Klassifikationsfehler jeweils auf den Trainings- und Testdaten (relative Häufigkeit von Fehlklassifikationen).

   (b) Zeichnen Sie die Kovariablen aller Patienten in ein Koordinatensystem und markieren Sie die Fälle gesund/krank (mit Farben) sowie Training-/Test-Datum (mit dunkel/hell).

   (c) Führen Sie eine lineare Diskriminanzanalyse für die Trainingsdaten durch, zeichnen Sie die erhaltene Entscheidungsgrenze in das Koordinatensystem und bestimmen Sie den Klassifikationsfehler jeweils auf den Trainings- und Testdaten (relative Häufigkeit von Fehlklassifikationen).

Abgabe am Dienstag, 17.1.23, über *Moodle*.

Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

11. Übungsblatt

1. Beweisen oder widerlegen Sie die Aussage, dass folgende Verteilungen Exponentialfamilien bilden. Bestimmen Sie gegebenenfalls den natürlichen Parameterraum.

   (a) Multinomialverteilung $(M(p_0, \dots, p_s; n))_{0<p_i<1, \sum_{i=1}^s p_i = 1}$;

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

   (c) Gleichmäßige Verteilung $(U([0, \vartheta]))_{\vartheta > 0}$.

2. Betrachten Sie die logistische Regression mit reellwertigen Kovariablen X_i, so dass $Y_i \,|\, X_i = x_i \sim \mathrm{Bern}((1 + e^{-(\beta_0 + \beta_1 x_i)})^{-1})$. Zeigen Sie, dass der MLE für $\beta = (\beta_0, \beta_1)^\top$ nicht existiert, falls eine exakte Trennung der Klassen in den Daten $(x_1, y_1), \dots, (x_n, y_n)$ möglich ist, das heißt $\xi \in \mathbb{R}$ existiert mit $y_i = \operatorname{sgn}(x_i - \xi)$ (bzw. $y_i = -\operatorname{sgn}(x_i - \xi)$).

   Verallgemeinern Sie dies für $\mathbb{R}^p$-wertige Kovariablen X_i, bei denen die Klassen durch eine linear-affine Hyperebene getrennt werden können.

3. Zeigen Sie:

   (a) Die Poissonverteilung mit Parameter $\lambda > 0$ bildet eine Exponentialfamilie in $T(k) = k$ mit natürlichem Parameter $\vartheta = \log \lambda \in \mathbb{R}$.

   (b) Betrachten Sie das GLM-Modell der Poissonregression mit $\log \lambda_i = \beta_0 + \beta_1 x_i$ für gegebene Designpunkte $x_1, \dots, x_n \in \mathbb{R}$ und $\beta = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$ unbekannt. Stellen Sie eine Gleichung für den MLE $\hat{\beta}$ auf und untersuchen Sie, ob der MLE existiert und eindeutig ist.

   Lektüretipp: In *Report 490 des Imperial College* werden *Growth, population distribution and immune escape of Omicron in England* mittels logistischer und Poisson-Regression untersucht.

Abgabe am Dienstag, 24.1.23, über *Moodle*.

Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

12. Übungsblatt

1. In der Vorlesung wurde $M_X := \sqrt{n/p}\max_{j=1,\dots,n}|\Pi_X e_j|$ definiert für die Orthogonalprojektion Π_X auf das Bild der Designmatrix $X \in \mathbb{R}^{n\times p}$ vom Rang p. Zeigen Sie:

   (a) $M_X \leqslant \sqrt{n/p}$.

   (b) Es gilt $\sum_{j=1}^n |\Pi_X e_j|^2 = p$, so dass $M_X \geqslant 1$ folgt.

   (c) Mit den Bezeichnungen $\Sigma_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ für $x_i = X^\top e_i$ sowie $\|X\|_{max} = \max_{i,j}|X_{ij}|$ gilt

   $$M_X \leqslant \frac{\|X\|_{max}}{\lambda_{min}(\Sigma_n)^{1/2}}.$$

2. Beweisen Sie für die Funktion $g_q(a) := -qa + \log(1+e^a)$, $a \in \mathbb{R}$, $q \in (0,1)$, die für $q = y$ und $a = \eta$ der Funktion $\ell_\eta(y)$ aus der logistischen Regression entspricht:

   (a) g_q ist strikt konvex.

   (b) Das globale Minimum von g_q liegt bei $a_q = \log(q/(1-q))$.

   (c) Für $|a - a_q| \leqslant 1$ gilt die Exzessbedingung

   $$g_q(a) \geqslant g_q(a_q) + \tfrac{q(1-q)}{2e}a^2.$$

   Tipp: Taylorentwicklung bis zum zweiten Glied.

   (d) Plotten Sie g_q und $a \mapsto g_q(a_q) + \frac{q(1-q)}{2e}a^2$ für $q \in \{0.1; 0.5; 0.9\}$.

3. Für jedes $\tau > 0$ gelte mit Wahrscheinlichkeit mindestens $1 - e^{-\tau}$ die Orakelungleichung

   $$\mathcal{E}_\ell(\hat\beta) \leqslant C \inf_{\beta\in\mathbb{R}^p} \mathcal{E}_\ell(\beta) + F(\tau)$$

   für eine Konstante $C \geqslant 1$ und eine deterministische, monoton wachsende Funktion $F : \mathbb{R}^+ \to \mathbb{R}^+$. Folgern Sie für den Erwartungswert die Orakelungleichung

   $$\mathbb{E}\left[\mathcal{E}_\ell(\hat\beta)\right] \leqslant C \inf_{\beta\in\mathbb{R}^p} \mathcal{E}_\ell(\beta) + \mathbb{E}\left[F(Z)\right]$$

   mit einer Exp(1)-verteilten Zufallsvariablen Z.

4. Lesen Sie Kapitel 2.1.2 *Sub-Gaussian Variables and Hoeffding bounds* im Buch *High-dimensional Statistics* von M. Wainwright und beweisen Sie die Hoeffding-Ungleichung für subgaußsche Zufallsvariablen (Prop. 2.5) im Detail.

Abgabe am Dienstag, 31.1.23, über *Moodle*.

Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

13. Übungsblatt

1. Betrachten Sie die bedingte Klassenwahrscheinlichkeit $\eta(x) = \mathbb{P}(Y = +1 \,|\, X = x)$, $x \in \mathbb{R}^p$, im LDA-Modell und überprüfen Sie, ob η für die kNN-Methode die verallgemeinerte Hölderbedingung

$$\forall x, y \in \mathbb{R}^p : \; |\eta(x) - \eta(y)| \leqslant L(x) G(x, |y - x|)^{1/\rho}$$

   mit $G(x,t) := \mathbb{P}(|X - x| \leqslant t)$ und geeigneten L und ρ erfüllt.

2. kNN-Klassifikation in der Praxis: Führen Sie für die Cleveland-Daten (vgl. Aufgabe 10.4) eine Klassifikation mit der kNN-Methode durch und testen Sie es analog zum Fall von LDA und logistischer Regression. Variieren Sie dabei $k \in \{1; 3; 10; 30\}$ und auch die Metrik $d(x,y) = (a(x_1 - y_1)^2 + b(x_2 - y_2)^2)^{1/2}$ mit $a, b \in \{1, 3\}$. Diskutieren Sie kurz die Ergebnisse und den Einfluss von k bzw. der Metrik.

Abgabe am Dienstag, 7.2.23, über *Moodle*.

Vorlesung *Methoden der Statistik*
Wintersemester 2022/23
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Eric Ziebell

14. Übungsblatt

1. Betrachten Sie die Funktion

$$J : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^+ \text{ mit } J(\beta, \beta_0) = \frac{1}{n} \sum_{i=1}^{n} \big(1 - Y_i(\langle X_i, \beta \rangle + \beta_0)\big)_+ + \tfrac{\lambda}{2}|\beta|^2$$

für gegebene $Y_i \in \{-1, +1\}$, $X_i \in \mathbb{R}^p$, $\lambda > 0$. Zeigen Sie:

   (a) Die Funktion $\beta \mapsto J(\beta, \beta_0)$ ist für jedes $\beta_0 \in \mathbb{R}$ strikt konvex und besitzt genau ein Minimum $\hat{\beta} = \hat{\beta}(\beta_0)$. J besitzt ein Minimum $(\hat{\beta}, \hat{\beta}_0)$ auf $\mathbb{R}^p \times \mathbb{R}$, das aber nicht notwendigerweise eindeutig ist.

   (b) Für beliebige Richtungsvektoren $v \in \mathbb{R}^p$ erhalten wir die einseitige Ableitung

$$\lim_{h \downarrow 0} \frac{J(\beta + hv, \beta_0) - J(\beta, \beta_0)}{h} = \lambda \langle \beta, v \rangle + \\ \frac{1}{n} \sum_{i=1}^{n} \Big(- Y_i \langle X_i, v \rangle \mathbf{1}\big(Y_i(\langle X_i, \beta \rangle + \beta_0) < 1\big) + (Y_i \langle X_i, v \rangle)_- \mathbf{1}\big(Y_i(\langle X_i, \beta \rangle + \beta_0) = 1\big)\Big).$$

   (c) Ist $v \in \mathbb{R}^p$ orthogonal zu allen Punkten X_i mit $Y_i(\langle X_i, \hat{\beta} \rangle + \beta_0) = 1$ für den Minimierer $\hat{\beta}$, so gilt

$$\langle \hat{\beta}, v \rangle = (\lambda n)^{-1} \sum_{i=1}^{n} Y_i \langle X_i, v \rangle \mathbf{1}\big(Y_i(\langle X_i, \hat{\beta} \rangle + \beta_0) < 1\big).$$

   (d) Es folgt die Existenz einer Darstellung

$$\hat{\beta} = \sum_{i=1}^{n} \alpha_i Y_i X_i$$

   mit $\alpha_i = (\lambda n)^{-1}$ im Fall $Y_i(\langle X_i, \hat{\beta} \rangle + \beta_0) < 1$, $\alpha_i = 0$ im Fall $Y_i(\langle X_i, \hat{\beta} \rangle + \beta_0) > 1$ und geeigneten α_i im Fall $Y_i(\langle X_i, \hat{\beta} \rangle + \beta_0) = 1$.

Alternativ: Verwenden Sie das Konzept der Subdifferentiale, um das Ergebnis in (d) herzuleiten.

2. Es sei $f \in L^1(\mathbb{R})$ eine reellwertige Funktion mit $f(x) \geqslant 0$ und $f(-x) = f(x)$ für alle $x \in \mathbb{R}$. Zeigen Sie, dass die Fouriertransformierte (oder charakteristische Funktion)

$$\varphi(u) = \int_{\mathbb{R}} f(x) e^{iux} dx, \quad u \in \mathbb{R},$$

reellwertig und symmetrisch ist. Folgern Sie ferner, dass $\sum_{i,j=1}^{m} \alpha_i \alpha_j \varphi(u_i - u_j) \geqslant 0$ für alle $m \in \mathbb{N}, \alpha_1, \ldots, \alpha_m, u_1, \ldots, u_m \in \mathbb{R}$ gilt.

Schließen Sie, dass der Gaußkern in der Tat positiv definit ist (*Tipp:* charakteristische Funktion der Normalverteilung). Welcher Kern ergibt sich analog aus $f(x) = e^{-|x|}$?

Abgabe am Dienstag, 14.2.23, über *Moodle*.