Dr. Stefan Richter
Statistische Analyse von maschinellen Lernalgorithmen
Wintersemester 2020/2021

5. Übungsblatt

Aufgabe 1 (Beispiel 4.1: Naiver Klassifizierer).

Gegeben seien i.i.d. Trainingsdaten (X_i, Y_i) eines Klassifizierungsproblems mit $\mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \{-1, +1\}$. Es sei $\tilde{L}(y, s) = (y - s)^2$. Wir nutzen außerdem die Abkürzung

$$\Delta := \{\delta_\beta(x) := x^T\beta : \beta \in \mathbb{R}^d\}.$$

Es wird folgender Algorithmus zur Klassifizierung vorgeschlagen:

$$\hat{\beta} :\in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \tilde{R}_n(\delta_\beta), \quad \tilde{R}_n(\delta) := \frac{1}{n}\sum_{i=1}^n \tilde{L}(Y_i, \delta(X_i)),$$

und $\hat{f}_n(x) := \operatorname{sign}(\hat{\delta}_n(x))$, wobei $\hat{\delta}_n(x) := \delta_{\hat{\beta}}(x) = x^T\hat{\beta}$. Anschaulich wird also auf (X_i, Y_i) eine lineare Regression angewandt.

(a) Zeigen Sie, dass $\tilde{L}(y, s) = \phi(-ys)$ mit geeignetem $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$.

(b) Ermitteln Sie mit Satz 3.19 eine Bayes-Regel δ^* für das Risiko $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$.

(c) Gilt die Kalibrierungsbedingung?

(d) Welche Modellannahme müsste gelten, damit $\delta^* \in \Delta$? Unter welcher Annahme an $\mathcal{X}$ kann dies überhaupt erfüllt sein?

(e) Ermitteln Sie eine Risikoübertragungsformel mit Hilfe von Satz 3.21.

(f) Es gelte nun die Rauschbedingung mit Parametern $q \geq 0$, $C > 0$. Geben Sie nun die Risikoübertragungsformel aus Satz 3.24 an, insbesondere im Fall $q = \infty$.

(g) Es sei $X \sim N(\mu, I_{d\times d})$ und $\eta(x) = \mathbb{P}(Y = 1|X = x) = \frac{1}{2} + \frac{1}{2}x^T\beta^*$. Es sei $\beta^* \in \mathbb{R}^d$ so gewählt, dass $\|\beta^*\|_2 = 2$ und $\mu^T\beta^* = 0$. Zeigen Sie, dass die Rauschbedingung mit $q = 1$ erfüllt ist.
Hinweis: Die Verteilungsfunktion $\Phi(t)$ der Standardnormalverteilung ist für $t \geq 0$ konkav, daher gilt $2\Phi(t) - 1 \leq 2\Phi'(0)t$.

Aufgabe 2 (Nachprüfen: Kalibrierungsbedingung und Risiko-Übertragungsformel).

Gegeben sei ein Klassifizierungsproblem mit $\mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \{-1, +1\}$. Es sei $\tilde{L}(y, s) = \phi(-ys)$, wobei wir zwei mögliche Funktionen $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ untersuchen:

- $\phi(x) = \phi_{hinge}(x) = \max\{1 + x, 0\}$, der so genannte hinge-Verlust,
- $\phi(x) = \phi_{exp}(x) = e^x$, der so genannte exponentielle Verlust.

(a) Überzeugen Sie sich, dass ϕ jeweils monoton wachsend und konvex ist und $\phi(0) = 1$ erfüllt.

(b) Ermitteln Sie mit Satz 3.19 eine Bayes-Regel δ^* für das Risiko $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$. *Hinweis für ϕ_{hinge}: Aus Monotoniegründen muss $z \mapsto \Phi_\eta(z)$ das Minimum in $z \in [-1, 1]$ annehmen.*

(c) Prüfen Sie, ob die Kalibrierungsbedingung erfüllt ist.

(d) Ermitteln Sie mit Satz 3.21 eine Risikoübertragungsformel.

Aufgabe 3 (Ermittlung des Wolfe-Duals und der Optimalitätsbedingungen der SVM).

Sei $C > 0$. Seien $\hat{\beta}_C$, $\hat{\beta}_{0,C}, \hat{\xi}$ Lösungen von

$$\min_{\beta \in \mathbb{R}^k, \beta_0 \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \qquad \text{unter NB} \qquad \forall i = 1, ..., n: \quad \begin{aligned} &Y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i, \\ &\xi_i \geq 0, \end{aligned}$$

Dieses Optimierungsproblem hat die Struktur

$$\min_{\theta \in \mathbb{R}^r} F(\theta) \quad \text{unter NB} \quad G(\theta) \leq 0.$$

mit $\theta = (\beta, \beta_0, \xi)$ und

$$F(\theta) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i, \quad G(\theta) = \begin{pmatrix} \big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big)_{i=1,\dots,n} \\ -\xi \end{pmatrix}.$$

Sei nun $p = (\alpha, \gamma) \in \mathbb{R}^{2n}_{\geq 0}$, und definiere die Lagrange-Funktion $L(\theta, p) = F(\theta) + G(\theta)^T p$.

(a) Zeigen Sie, dass $\nabla_\theta L(\theta, p) = 0$ äquivalent ist zu

$$\beta = \sum_{i=1}^n \alpha_i Y_i X_i, \qquad \sum_{i=1}^n \alpha_i Y_i = 0, \qquad \forall i: C - \alpha_i - \gamma_i = 0.$$

(b) Zeigen Sie, dass das Wolfe-Dual $\sup_{\theta \in \mathbb{R}^r, p \in \mathbb{R}^k_{\geq 0}} L(\theta, p)$ unter der NB $\nabla_\theta L(\theta, p) = 0$ äquivalent ist zum Optimierungsproblem

$$\min_{\alpha \in \mathbb{R}^n} \left\{\frac{1}{2}\alpha^T Q\alpha - \mathbb{1}^T\alpha\right\} \qquad \text{unter NB} \qquad \mathbb{Y}^T\alpha = 0, \quad 0 \leq \alpha \leq C,$$

wobei $\mathbb{1} = (1, ..., 1)^T$, $\mathbb{Y} = (Y_1, ..., Y_n)^T$ und $Q := (Q_{ij})_{i,j=1,...,n}$ mit $Q_{ij} = Y_iY_jX_i^TX_j$.

(c) Folgern Sie aus den Optimalitätsbedingungen $\nabla_\theta L(\hat{\theta}, \hat{p}) = 0$, $G(\hat{\theta})^T\hat{p} = 0$, $\hat{p} \geq 0$ und $G(\hat{\theta}) \leq 0$, dass folgender Zusammenhang zwischen den Lösungen des ursprünglichen Optimierungsproblems und den Lösungen $\hat{\alpha}$ des Wolfe-Duals besteht:

$$\hat{\beta}_C = \sum_{i=1}^n \hat{\alpha}_i Y_i X_i, \qquad \hat{\beta}_{0,C} = Y_i - X_i^T\hat{\beta}_C \quad \text{mit einem } i \text{ mit } 0 < \hat{\alpha}_i < C.$$

Homepage der Vorlesung:
`https://ssp.math.uni-heidelberg.de/sam-ws2020/`