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

5. Übungsblatt

Aufgabe 1 (Beispiel 4.1: Naiver Klassifizierer)
Lösung:

(a) Es gilt wegen $y \in \{+1, -1\}$:

$$\tilde{L}(y, s) = (y - s)^2 = y^2 - 2ys + s^2 = 1 - 2ys + (ys)^2 = (1 - ys)^2 = \phi(-ys)$$

mit $\phi(x) = (1 + x)^2$.

(b) Es ist $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1 - \eta) = (1 - z)^2\eta + (1 + z)^2(1 - \eta)$.
Suche Minimierer von $z \mapsto \Phi_\eta(z)$:

$$0 = \Phi'_\eta(z) = -2(1 - z)\eta + 2(1 + z)(1 - \eta) = 2 + 2z - 4\eta \quad \Rightarrow \quad z = 2\eta - 1.$$

Satz 3.19 $\Rightarrow \delta^*(x) = 2\eta(x) - 1$ mit $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

(c) Es gilt $\delta^*(x) > 0 \iff \eta(x) > \frac{1}{2} \iff f^*(x) = \text{argmax}_{k\in\{+1,-1\}} \mathbb{P}(Y = k|X = x) = 1$.
$\Rightarrow f^*(x) = \text{sign}(\delta^*(x))$
$\Rightarrow$ Kalibrierungsbedingung ist erfüllt.

(d) $\delta^* \in \Delta$ ist äquivalent zu $\forall x \in \mathcal{X} : 2\eta(x) - 1 = \delta^*(x) = x^T\beta^*$ für ein $\beta^* \in \mathbb{R}^d$, d.h. es müsste gelten

$$\eta(x) = \frac{1}{2} + \frac{1}{2}x^T\beta^*.$$

Da die linke Seite $\in [0, 1]$ ist, kann dies nur gelten, wenn $\mathcal{X}$ beschränkt ist.

(e) Schon in (b) gesehen: $g(\eta) = \text{argmin}_{z\in\mathbb{R}} \Phi_\eta(z) = 2\eta - 1$.
Außerdem

$$H(\eta) = \Phi_\eta(g(\eta)) = (2 - 2\eta)^2\eta + (2\eta)^2(1 - \eta) = 4\eta(1 - \eta).$$

Damit ist

$$1 - H(\eta) = 1 - 4\eta + 4\eta^2 = 4(\eta - \frac{1}{2})^2 \quad \Rightarrow \quad (\frac{1}{2})^2(1 - H(\eta)) = (\eta - \frac{1}{2})^2.$$

Außerdem $\phi(0) = 1$, ϕ konvex. Satz 3.21 $\Rightarrow$ Mit $s = 2$, $C_H = \frac{1}{2}$ gilt die Risikoübertragungsformel mit

$$G(r) = 2C_H r^{1/s} = r^{1/2}.$$

(f) Fall $q \in [0, \infty)$: $G(r) = 4C_H^{\frac{s(q+1)}{q+s}} C^{-\frac{1}{q+s}} r^{\frac{q+1}{q+s}} = 4(\frac{1}{2})^{\frac{2(q+1)}{q+2}} C^{-\frac{1}{q+2}} r^{\frac{q+1}{q+2}}$.
Fall $q = \infty$: $G(r) = 2\frac{C_H^s}{c^{s-1}} r = \frac{1}{2c} \cdot r$.

(g) Es gilt wegen $\frac{1}{2}X^T\beta^* \sim N(\frac{1}{2}\mu^T\beta^*, \frac{1}{4}\|\beta^*\|_2^2) = N(0,1)$:

$$\begin{aligned}\mathbb{P}(|\eta(X) - \frac{1}{2}| \leq t) &= \mathbb{P}(\frac{1}{2}|X^T\beta^*| \leq t) \overset{Z\sim N(0,1)}{=} \mathbb{P}(|Z| \leq t) = 2\Phi(t) - 1 \\ &\overset{\text{Hinweis}}{\leq} 2\Phi'(0)t = \sqrt{\frac{2}{\pi}} \cdot t.\end{aligned}$$

Hier ist $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)dy$ und damit $\Phi'(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$.
Damit ist die Rauschbedingung

$$\mathbb{P}(|\eta(X) - \frac{1}{2}| \leq t) \leq Ct^q \quad \forall t > 0$$

erfüllt mit $C = \sqrt{\frac{2}{\pi}}$, $q = 1$.

Aufgabe 2 (Nachprüfen: Kalibrierungsbedingung und Risiko-Übertragungsformel)
Lösung: Für $\phi(x) = e^x$:

(a) $\phi'(x) = e^x > 0 \Rightarrow \phi$ monoton wachsend,
$\phi''(x) > 0 \Rightarrow \phi$ konvex,
$\phi(0) = e^0 = 1$.

(b) Es ist $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1-\eta) = e^{-z}\eta + e^z(1-\eta)$. Suche Minimierer von $z \mapsto \Phi_\eta(z)$:

$$0 = \Phi_\eta'(z) = -e^{-z}\eta + e^z(1-\eta) \quad \Rightarrow \quad z = \frac{1}{2}\log(\frac{\eta}{1-\eta}).$$

Satz 3.19 $\Rightarrow \delta^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)})$ mit $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

(c) Es gilt $\delta^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)}) > 0 \Longleftrightarrow \frac{\eta(x)}{1-\eta(x)} > 1 \Longleftrightarrow \eta(x) > \frac{1}{2} \Longleftrightarrow f^*(x) = \text{argmax}_{k\in\{+1,-1\}} \mathbb{P}(Y = k|X = x) = 1$.
$\Rightarrow f^*(x) = \text{sign}(\delta^*(x))$
$\Rightarrow$ Kalibrierungsbedingung ist erfüllt.

(d) Schon in (b) gesehen: $g(\eta) = \text{argmin}_{z\in\mathbb{R}} \Phi_\eta(z) = \frac{1}{2}\log(\frac{\eta}{1-\eta})$.
Außerdem

$$\begin{aligned}H(\eta) &= \Phi_\eta(g(\eta)) = \exp\Big(-\frac{1}{2}\log(\frac{\eta}{1-\eta})\Big)\eta + \exp\Big(\frac{1}{2}\log(\frac{\eta}{1-\eta})\Big)(1-\eta) \\ &= \Big(\frac{\eta}{1-\eta}\Big)^{-1/2}\eta + \Big(\frac{\eta}{1-\eta}\Big)^{1/2}(1-\eta) \\ &= 2(1-\eta)^{1/2}\eta^{1/2} = 2((1-\eta)\eta)^{1/2}.\end{aligned}$$

Damit ist

$$p(\eta) := 1 - H(\eta) = 1 - 2((1-\eta)\eta)^{1/2}.$$

Es gilt $p'(\eta) = 2(\eta - \frac{1}{2}) \cdot (\eta(1-\eta))^{-1/2}$, $p''(\eta) = \frac{1}{2}(\eta(1-\eta))^{-3/2} \geq \frac{1}{2}(\frac{1}{4})^{-3/2} = 4$.
Taylor-Entwicklung um $\eta = \frac{1}{2}$:

$$1 - H(\eta) = p(\eta) = p(\frac{1}{2}) + (\eta - \frac{1}{2})p'(\frac{1}{2}) + \frac{1}{2}(\eta - \frac{1}{2})^2 p''(\xi) \geq \frac{1}{2}(\eta - \frac{1}{2})^2 \cdot 4 = 2(\eta - \frac{1}{2})^2.$$

Außerdem $\phi(0) = 1$, ϕ konvex. Satz 3.21 $\Rightarrow$ Mit $s = 2$, $C_H = \frac{1}{\sqrt{2}}$ gilt die Risikoübertragungsformel mit

$$G(r) = 2C_H r^{1/s} = \sqrt{2}r^{1/2}.$$

Für $\phi(x) = \max\{1 + x, 0\}$:

(a) $1+x, 0$ monoton wachsend $\Rightarrow \phi(x) = \max\{1+x, 0\}$ monoton wachsend (Maximum erhält Monotonie).
$1 + x, 0$ konvex $\Rightarrow \phi(x) = \max\{1 + x, 0\}$ konvex (Maximum erhält Konvexität).
$\phi(0) = \max\{1, 0\} = 1$.

(b) Es ist $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1 - \eta) = \max\{1 - z, 0\}\eta + \max\{1 + z, 0\}(1 - \eta)$. Suche Minimierer von $z \mapsto \Phi_\eta(z)$.
Für $z < -1$ ist $\Phi_\eta(z) = (1 - z)\eta$ monoton fallend, für $z > 1$ ist $\Phi_\eta(z) = (1 + z)(1 - \eta)$ monoton wachsend. $\Rightarrow$ Minimierer liegt in $[-1, 1]$.
Für $z \in [-1, 1]$ ist

$$\Phi_\eta(z) = (1 - z)\eta + (1 + z)(1 - \eta) = 1 + 2z \cdot (\frac{1}{2} - \eta).$$

Für $\eta > \frac{1}{2}$ wird dies minimal für $z = 1$, für $\eta < \frac{1}{2}$ wird dies minimal für $z = -1$. Für $\eta = \frac{1}{2}$ gibt es keinen eindeutigen Minimierer.
$\Rightarrow z = \mathrm{sign}(\eta - \frac{1}{2})$ ist Minimierer
Satz 3.19 $\Rightarrow \delta^*(x) = \mathrm{sign}(\eta(x) - \frac{1}{2})$.

(c) Es gilt $\delta^*(x) = \mathrm{sign}(\eta(x) - \frac{1}{2}) = f^*(x)$, insbesondere also $\Rightarrow f^*(x) = \mathrm{sign}(\delta^*(x))$
$\Rightarrow$ Kalibrierungsbedingung ist erfüllt.
Besonderheit hier: $\delta^*(x)$ selbst ist auch diskret mit Werten in $\{-1, +1\}$!

(d) Schon in (b) gesehen: $g(\eta) = \mathrm{sign}(\eta - \frac{1}{2})$.
Außerdem

$$H(\eta) \quad = \quad \Phi_\eta(g(\eta)) = 1 + 2\mathrm{sign}(\eta - \frac{1}{2}) \cdot (\frac{1}{2} - \eta) = 1 - 2|\eta - \frac{1}{2}|.$$

Damit ist

$$1 - H(\eta) = 2|\eta - \frac{1}{2}|.$$

Außerdem $\phi(0) = 1$, ϕ konvex. Satz 3.21 $\Rightarrow$ Mit $s = 1$, $C_H = \frac{1}{2}$ gilt die Risikoübertragungsformel mit

$$G(r) = 2C_H r^{1/s} = r.$$

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

(a) Es ist

$$L(\theta, p) \quad = \quad \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \quad + \sum_{i=1}^n \alpha_i\big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big) \quad - \sum_{i=1}^n \gamma_i\xi_i$$

Untersuche die drei Ableitungen: $\nabla_\theta = (\nabla_\beta, \nabla_{\beta_0}, \nabla_\xi)$:

$$\nabla_\beta L(\theta, p) = \beta - \sum_{i=1}^n \alpha_i Y_i X_i, \qquad \nabla_{\beta_0} L(\theta, p) = \sum_{i=1}^n \alpha_i Y_i, \qquad \nabla_{\xi_i} L(\theta, p) = C - \alpha_i - \gamma_i.$$

Damit folgt die Aussage, denn die Gleichungen $\nabla_\theta L(\theta, p) = 0$ ergeben genau die angegebenen Gleichungen.

(b) Wir setzen die Nebenbedingungen $\nabla_\theta L(\theta, p) = 0$ in die Optimierungsfunktion $L(\theta, p)$ ein und formen um. Dabei achten wir darauf, dass eventuelle Nebenbedingungen für α mitgenommen werden. Es gilt

$$\begin{aligned}
L(\theta, p) &= \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i\big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big) - \sum_{i=1}^n \gamma_i\xi_i \\
&= \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n \xi_i \cdot \underbrace{(C - \alpha_i - \gamma_i)}_{=0} - \beta_0 \underbrace{\sum_{i=1}^n \alpha_i Y_i}_{=0} - \underbrace{\Big(\sum_{i=1}^n \alpha_i Y_i X_i^T\Big)}_{=\beta^T}\beta \\
&\overset{\nabla_\theta L(\theta,p)=0}{=} \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n \alpha_i - \|\beta\|_2^2 \\
&= \sum_{i=1}^n \alpha_i - \frac{1}{2}\|\beta\|_2^2 \\
&= \sum_{i=1}^n \alpha_i - \frac{1}{2}\Big\|\sum_{i=1}^n \alpha_i Y_i X_i\Big\|_2^2 \\
&= \underbrace{\sum_{i=1}^n \alpha_i}_{=\mathbb{1}^T\alpha} - \underbrace{\frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j X_i^T X_j Y_i Y_j}_{=\alpha^T Q\alpha} \\
&= \mathbb{1}^T\alpha - \frac{1}{2}\alpha^T Q\alpha.
\end{aligned}$$

Maximierer von $L(\theta, p)$ ist Minimierer von $-L(\theta, p)$. Dies liefert die angegebene Optimierungsfunktion.
Nebenbedingungen an α:

- $\nabla_{\beta_0} L(\theta, p) = 0$ ist äquivalent zu $\mathbb{Y}^T\alpha = \sum_{i=1}^n \alpha_i Y_i = 0$.
- $\nabla_\beta L(\theta, p) = \beta - \sum_{i=1}^n \alpha_i Y_i X_i = 0$ ist irrelevant, da β nicht mehr in der Optimierungsfunktion vorkommt.
- $\nabla_{\xi_i} L(\theta, p) = C - \alpha_i - \gamma_i = 0$ liefert $\alpha_i = C - \gamma_i \leq C$, da $\gamma_i \geq 0$. Daher erhält man die Nebenbedingung $\alpha_i \leq C$.

(c) Optimalitätsbedingung $\nabla_\beta L(\hat{\theta}, \hat{p}) = 0$ liefert $\hat{\beta} = \sum_{i=1}^n \alpha_i Y_i X_i$.
Wegen $G(\hat{\theta}) \leq 0$, $\hat{p} \geq 0$ ist

$$0 = G(\hat{\theta})^T\hat{p} = \sum_j G(\hat{\theta})_j \hat{p}_j \quad \Longleftrightarrow \quad \forall j : G(\hat{\theta})_j\hat{p}_j = 0.$$

Insbesondere gilt für alle $i = 1, ..., n$:

$$\hat{\gamma}_i\hat{\xi}_i = 0, \qquad \hat{\alpha}_i(1 - \hat{\xi}_i - Y_i(X_i^T\hat{\beta} + \hat{\beta}_0)) = 0. \qquad (*)$$

Es sei $i \in \{1, ..., n\}$ mit $\hat{\alpha}_i \in (0, C)$.
$\Rightarrow 0 = \nabla_{\beta_0} L(\hat{\theta}, \hat{p}) = C - \hat{\alpha}_i - \hat{\gamma}_i \Rightarrow \hat{\gamma}_i = C - \hat{\alpha}_i > 0$.
(*) $\Rightarrow \hat{\xi}_i = 0$
(*), $\hat{\alpha}_i \neq 0 \Rightarrow 0 = 1 - \hat{\xi}_i - Y_i(X_i^T\hat{\beta} + \hat{\beta}_0) = 1 - Y_i(X_i^T\hat{\beta} + \hat{\beta}_0) \Rightarrow 0 = Y_i - Y_i^2(X_i^T\hat{\beta} + \hat{\beta}_0) = Y_i - X_i^T\hat{\beta} - \hat{\beta}_0$
$\Rightarrow \hat{\beta}_0 = Y_i - X_i^T\hat{\beta}$.

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