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

1. Übungsblatt

Aufgabe 1: Diskussion: Beweis von Satz 2.5
Lösung:

(a) Markov-Ungleichung:

$$\mathbb{P}(\|A\mathbb{e}\|_2 \geq \sigma\|A\|_F\sqrt{t}) \leq \frac{\mathbb{E}[\|A\mathbb{e}\|_2^2]}{(\sigma\|A\|_F\sqrt{t})^2}.$$

Hier ist

$$\mathbb{E}[\|A\mathbb{e}\|_2^2] = \mathbb{E}[\mathrm{tr}(A\mathbb{e}\mathbb{e}^T A^T)] = \mathrm{tr}(A\mathbb{E}[\mathbb{e}\mathbb{e}^T]A^T) = \mathrm{tr}(A\sigma^2 I_{d\times d}A^T) = \sigma^2\mathrm{tr}(AA^T) = \sigma^2\|A\|_F^2.$$

Einsetzen oben:

$$\frac{\mathbb{E}[\|A\mathbb{e}\|_2^2]}{(\sigma\|A\|_F\sqrt{t})^2} = \frac{1}{t}.$$

(b) Abgesehen von der Konstante c_2 sind die Ausdrücke in der Wahrscheinlichkeit dieselben für $t \geq 1$.
Das Lemma der Vorlesung ist aber eine wesentlich bessere Abschätzung, da für große t der Term e^{-t} viel schneller fällt als $\frac{1}{t}$. Das Lemma liefert also eine viel bessere Abschätzung für die Wahrscheinlichkeit des Ereignisses (allerdings auch unter stärkeren Voraussetzungen, für die Rechnung in (a) brauchte man nur Existenz 2. Moment von $\mathbb{e}$, keine Normalverteilung).

(c) Wegen $\|\hat{\Sigma} - \Sigma\| \leq \|\hat{\Sigma} - \Sigma\|_F$ gilt mit Markov-Ungleichung:

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\| \geq x) \leq \frac{\mathbb{E}[\|\hat{\Sigma} - \Sigma\|_F^2]}{x^2}.$$

Es gilt

$$\begin{aligned}
\mathbb{E}[\|\hat{\Sigma}-\Sigma\|_F^2] &= \sum_{j,k=1}^{d}\mathbb{E}\big[(\hat{\Sigma}_{jk}-\Sigma_{jk})^2\big] \\
&= \sum_{j,k=1}^{d}\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{ik}-\mathbb{E}[X_{ij}X_{ik}]\Big)^2\Big] \\
&= \sum_{j,k=1}^{d}\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{ik}\Big) \\
&= \sum_{j,k=1}^{d}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_{ij}X_{ik}) \\
&= \sum_{j,k=1}^{d}\frac{1}{n}\big\{\mathbb{E}[X_{1j}^2X_{1k}^2]-\mathbb{E}[X_{1j}X_{1k}]\big\} \\
&\overset{\text{Hinweis}}{=} \frac{1}{n}\sum_{j,k=1}^{d}\big\{\Sigma_{jj}\Sigma_{kk}+2\Sigma_{jk}^2-\Sigma_{jk}\big\} \\
&= \frac{1}{n}\Big\{\sum_{j,k=1}^{d}\Sigma_{jj}\Sigma_{kk}+\sum_{j,k=1}^{d}\Sigma_{jk}^2\Big\} \\
&= \frac{1}{n}\big\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\big\}.
\end{aligned}$$

Das liefert die Aussage.

(d) Substitution: $x' := \frac{1}{\sqrt{n}}\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}^{1/2}x$. Dann gilt mit (c):

$$\begin{aligned}
&\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq \frac{1}{\sqrt{n}}\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}^{1/2}x\Big) \\
\leq\ &\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq x'\Big) \leq \frac{1}{n}\frac{\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}}{(x')^2} = \frac{1}{x}.
\end{aligned}$$

(e) Falls $d \leq n$ und $n \geq x \geq d$ (das nehmen wir zum Vergleich einfach an), so ist die Aussage vom Lemma

$$\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq c_1\|\Sigma\|\cdot\sqrt{\frac{x}{n}}\Big) \leq e^{-x}$$

Die Aussage aus (d) ist jedoch

$$\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq \{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}^{1/2}\sqrt{\frac{x}{n}}\Big) \leq \frac{1}{x}.$$

Hier gibt es also zwei Nachteile gegenüber der Aussage aus dem Lemma: Zunächst ist wieder $\frac{1}{x}$ für große x eine wesentlich schlechtere Abschätzung als e^{-x}. Außerdem ist der Vergleichwert $\|\Sigma\|$ in dem Lemma wesentlich kleiner als $\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}^{1/2}$ in der Ungleichung aus (d). Zum Vergleich: Ist $\Sigma = I_{d\times d}$, so ist $\|\Sigma\| = 1$ aber $\{\mathrm{tr}(\Sigma)^2+\|\Sigma\|_F^2\}^{1/2} = \sqrt{2d}$. Das bedeutet, das Lemma macht auch hier eine schärfere Aussage.

Aufgabe 2: Ridge-Schätzer
Lösung:

(a) Jede Komponente β_j von β entspricht dem Einfluss des zugehörigen X_j. Die Größe von β_j drückt also aus, wie stark der Wert von X_j in Y eingeht. In dem modifizierten Modell kodiert β_1 aber nur den Erwartungswert von Y (ohne irgendwelche Einflüsse von X, denn X_1 wird ja konstant auf 1 gesetzt).
Der Sinn der Bestrafung von β ist es, dass nur die für Y wichtigen Komponenten von X ausgewählt werden. Ist der Erwartungswert von Y aber nicht null, so ist dieser immer wichtig. Der Erwartungswert und seine Größe (im Modell dann gegeben durch β_1, falls $\mathbb{E}X = 0$), sollte daher nicht bestraft werden.

(b) Es gilt dann für die Optimierungsfunktion (bezeichnet hier mit F):

$$F(\beta) = \hat{R}_n(\beta) + \lambda \cdot \sum_{j=2}^{d} \beta_j^2 = \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \sum_{j=2}^{d} \beta_j^2$$

Ableiten und Nullsetzen (mit $E = \mathrm{diag}(0, 1, ..., 1) \in \mathbb{R}^{d\times d}$):

$$0 = \nabla_\beta F(\beta) = -\frac{2}{n}\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\beta) + 2\lambda E\beta = -\frac{2}{n}\mathbb{X}^T\mathbb{Y} + 2(\frac{1}{n}\mathbb{X}^T\mathbb{X} + \lambda E)\beta.$$

Damit $\hat{\beta} = (\mathbb{X}^T\mathbb{X} + n\lambda E)^{-1}\mathbb{X}^T\mathbb{Y}$.

Aufgabe 3: Große Abweichungen
Lösung:

(a) Markov-Ungleichung:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}e^{cX}}{e^{ct}} = e^{\frac{c^2}{2} - ct}.$$

Hierbei haben wir genutzt, dass für $X \sim N(0, 1)$ gilt: $\mathbb{E}e^{cX} = e^{\frac{c^2}{2}}$.
Einsetzen von $c = t$:

$$\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2}}.$$

(b) Hier nutzen wir $\frac{X}{\sigma} \sim N(0, 1)$. Mit (a) folgt

$$\mathbb{P}(|X| > \sigma t) \leq \mathbb{P}(X > \sigma t) + \mathbb{P}(-X > \sigma t) = \mathbb{P}(\frac{X}{\sigma} > t) + \mathbb{P}(-\frac{X}{\sigma} > t) \leq 2e^{-\frac{t^2}{2}}.$$

Mit $t' := \sqrt{2t}$ folgt:

$$\mathbb{P}(|X| > \sigma\sqrt{2t}) = \mathbb{P}(|X| > \sigma t') \leq 2e^{-\frac{(t')^2}{2}} = 2e^{-t}.$$

(c) Die Aussage in (a) lautet für $X \sim N(0, \sigma^2)$: $\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$.
Nach dem zentralen Grenzwertsatz erwartet man für $n \to \infty$: $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \xrightarrow{D} N(0, \sigma^2)$ und daher

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\Big) \to \mathbb{P}(X > t) \overset{s.o.}{\leq} e^{-\frac{t^2}{2\sigma^2}}.$$

Die Bernstein-Ungleichung liefert hier eine zugehörige nicht-asymptotische Aussage (also ohne $n \to \infty$) mit

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\Big) \leq \exp\Big(-\frac{1}{2}\frac{t^2}{\sigma^2 + \frac{Mt}{\sqrt{n}}}\Big).$$

Man kann sehen, dass obiger Ausdruck für $n \to \infty$ wieder gegen $e^{-\frac{t^2}{2\sigma^2}}$ konvergiert. In diesem Sinne ist die Bernstein-Ungleichung also 'optimal', da sie eine Abschätzung für endliches n einer Summe von beliebigen i.i.d. Zufallsvariablen X_i liefert, die für $n \to \infty$ gegen die 'unvermeidliche' Abschätzung einer Normalverteilung konvergiert. Wenn man die Verteilung der X_i nicht kennt, liefert die Bernstein-Ungleichung also eine sehr gute Abschätzung, besonders für große n.

(d) Wegen $\frac{a}{b+c} \geq \frac{1}{2}\min\{\frac{a}{b}, \frac{a}{c}\}$ für $a, b, c > 0$ gilt

$$\exp\Big(-\frac{1}{2}\frac{x^2}{\sigma^2 + \frac{Mx}{\sqrt{n}}}\Big) \leq \exp\Big(-\frac{1}{4}\min\{\frac{x^2}{\sigma^2}, \frac{x}{M/\sqrt{n}}\}\Big).$$

Es folgt

$$\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \sqrt{n}\sigma\sqrt{t} + Mt\Big) &= \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \underbrace{\sigma\sqrt{t} + \frac{M}{\sqrt{n}}t}_{=:x'}\Big) \\
&\leq \exp\Big(-\frac{1}{4}\min\{\frac{(x')^2}{\sigma^2}, \frac{x'}{M/\sqrt{n}}\}\Big) \\
&\leq e^{-\frac{t}{4}}.
\end{aligned}$$

Die letzte Ungleichung gilt wegen $(x')^2 \geq t\sigma^2$ (nutze nur 1. Summand von x') und $x' \geq \frac{M}{\sqrt{n}}t$ (nutze nur 2. Summand von x').

(e) Wähle $t = 4\log(\frac{1}{\delta})$. Dann gilt mit (d):

$$\mathbb{P}\Big(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq 2\sqrt{n}\sigma\sqrt{\log(\frac{1}{\delta})} + 4M\log(\frac{1}{\delta})\Big) \leq e^{-t/4} = \delta.$$

Gegenereignis: Mit Wahrscheinlichkeit $\geq 1 - \delta$ gilt

$$\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq 2\sqrt{n}\sigma\sqrt{\log(\frac{1}{\delta})} + 4M\log(\frac{1}{\delta})$$

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