GRUNDLAGEN DES MASCHINELLEN LERNENS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Marvin Häußermann (marvin.haeussermann@uni-tuebingen.de)
Tobias Lang (t.lang@uni-tuebingen.de)
Mathias Schickel (mathias.schickel@uni-tuebingen.de)

Andreas Schilling
Sommersemester 2020

# Blatt 3

Ausgabe: 29.05.2020; Abgabe: bis 12.06.2020, 09:00 Uhr.

*In der Mathematik versteht man die Dinge nicht. Man gewöhnt sich nur an sie.*
John von Neumann

## Organisatorisches und Formales

- Es gibt *genau ein* Gruppenmitglied der *Vierer*gruppe die Lösung in Form eines Archivs ab, am besten im `zip`-Format.
- Das Archiv und der darin enthaltene Ordner mit allen Dateien, die zur Abgabe gehören, sowie das dort mitgelieferte PDF(!) mit der Textabgabe müssen der Namenskonvention `nachname1-nachname2-nachname3-nachname4-uebungX` entsprechen. Sonderzeichen sollten dabei nicht verwendet werden und die Namen sollen alphabetisch sortiert sein, z. B. `maier-mueller-schmidt-schulz-uebung2.zip`. (Die Zahl der angegebenen Namen soll dabei der Größe der Gruppe entsprechen.)
- Auf gute und prägnante Formulierung sowie Rechtschreibung und Zeichensetzung ist zu achten. Ausschweifende Antworten sowie mangelnde sprachliche Sorgfalt können mit Punktabzug geahndet werden.
- Textabgaben im Code werden **nicht gewertet**.
- Abbildungen dürfen **nicht übereinander geplottet** werden, außer es wird explizit verlangt.
- Insofern Dateien mit Funktionsrümpfen zur Verfügung gestellt wurden, sind diese beim Verfassen des Programmcodes zu verwenden und **in der vorgegebenen Ordnerstruktur** wieder abzugeben. Die Form der Ein- und Ausgabeparameter darf dabei **nicht verändert werden**.

**Aufgabe 1** (Grundlagen der Parameterschätzung)

a) In Anwendungen ist die Likelihoodfunktion im Allgemeinen nicht bekannt. Sie muss daher aus statistisch erhobenen Daten $D = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ geschätzt werden. Dazu wird oftmals vorausgesetzt, dass die likelihood durch eine parametrisierte Wahrscheinlichkeitsdichte $p(\cdot \mid \boldsymbol{\vartheta})$ mit Parametervektor $\boldsymbol{\vartheta}$ gegeben ist, deren prinzipielle Form bekannt ist, sodass nur die Parameter geschätzt werden müssen. Zusätzlich nimmt man oftmals an, dass die erhobenen Daten stochastisch unabhängig sind. Damit ergibt sich

$$p(D \mid \boldsymbol{\vartheta}) = \prod_{k=1}^{n} p(\boldsymbol{x}_k \mid \boldsymbol{\vartheta}). \tag{1}$$

Diesen Term möchte man nun über $\boldsymbol{\vartheta}$ maximieren. Das bedeutet, dass man denjenigen Parametervektor $\boldsymbol{\vartheta}^*$ bestimmen möchte, der die beobachteten Daten $D$ am wahrscheinlichsten macht, d. h.

$$\boldsymbol{\vartheta}^* := \arg\max_{\boldsymbol{\vartheta}} \prod_{k=1}^{n} p(\boldsymbol{x}_k \mid \boldsymbol{\vartheta}).$$

Warum verwendet man bei der Suche nach den Maximalstellen der Wahrscheinlichkeitsdichte $p(D \mid \boldsymbol{\vartheta})$ (betrachtet als Funktion von $\boldsymbol{\vartheta}$) ihren Logarithmus $\ell(\boldsymbol{\vartheta})$ (bezeichnet als *log-likelihood*) und wieso ist dies erlaubt? Welche analytisch einfacher zu handhabende Darstellung erhält man für $\ell(\boldsymbol{\vartheta})$, also für den Logarithmus von Gleichung (1)? Nenne die Eigenschaften des Logarithmus, die dessen Anwendung hier einerseits gestatten und andererseits günstig machen.

b) Welche Ausdrücke erhält man für die Parameter $\mu$ und $\sigma$ für eine univariate Normalverteilung bei Maximum-Likelihood-Schätzung?

c) Nenne mindestens zwei wichtige Annahmen der Bayesschen Parameterschätzung.

d) Nenne und beschreibe kurz die drei wesentlichen Arten von Fehlern, die bei der Parameterschätzung und der Klassifikation auf der Basis der geschätzten Dichten auftreten können.

**Hinweis:** Erinnert Euch für das Lösen von Aufgabenteil a) daran, wie man zweimal differenzierbare reellwertige Funktionen $f : \mathbb{R}^n \to \mathbb{R}$ maximieren kann und welche Bedingungen in Maximalstellen $\boldsymbol{\vartheta} \in \mathbb{R}$ gelten müssen (für diese Werte ist $f(\boldsymbol{\vartheta})$ ein lokales Maximum von $f$). Dies ist auch im Zusammenhang von Aufgabe 2 hilfreich.

**Aufgabe 2** (Maximum-Likelihood-Schätzung)

Gegeben seien Ergebnisse $x_1, \dots, x_n$ von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \dots, X_n$, wobei $X_i \sim \mathcal{N}(\mu, \sigma^2)$ für jedes $i = 1, \dots, n$ gelte. Die Wahrscheinlichkeitsdichte der Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist gegeben durch

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Erkläre den folgenden Beweis, dass für die *Maximum-Likelihood-Schätzer* $\widehat{\mu}$ und $\hat{\sigma}^2$

$$\widehat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{und} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_k - \widehat{\mu})^2$$

gilt. Welche Größen bezeichnen diese Ausdrücke inhaltlich?

**Beweis.** Die Dichte der Normalverteilung hat die Parameter $\vartheta_1 = \mu$ und $\vartheta_2 = \sigma^2$. Daher gilt

$$\nabla \ell(\boldsymbol{\vartheta}) = \sum_{k=1}^{n} \nabla \log p(x_k \mid \boldsymbol{\vartheta}) = \sum_{k=1}^{n} \begin{pmatrix} \frac{\partial}{\partial \vartheta_1} \log p(x_k \mid \boldsymbol{\vartheta}) \\ \frac{\partial}{\partial \vartheta_2} \log p(x_k \mid \boldsymbol{\vartheta}) \end{pmatrix} = \sum_{k=1}^{n} \begin{pmatrix} \frac{1}{\vartheta_2}(x_k - \vartheta_1) \\ -\frac{1}{2\vartheta_2} + \frac{1}{2\vartheta_2^2}(x_k - \vartheta_1)^2 \end{pmatrix}.$$

Daraus folgt

$$\hat{\vartheta}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$$

und des Weiteren

$$\frac{1}{2\hat{\vartheta}_2^2} \cdot \left(\sum_{k=1}^{n}(x_k - \hat{\vartheta}_1)^2\right) = \frac{n}{2\hat{\vartheta}_2}$$

und daher

$$\hat{\vartheta}_2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\vartheta}_1)^2.$$

◆

Bei der Erläuterung des Beweises sollen die grundlegende Strategie (warum beweist das, dass die finalen Ausdrücke tatsächlich die Maximum-Likelihood-Schätzer der Parameter der Normalverteilung sind) und außerdem die einzelnen aufgeführten Schritte erklärt werden. Dabei soll die Herkunft jeder Formel deutlich werden und ebenso der Grund, warum sie betrachtet wird. Vor jedem im obigen Beweis aufgeführten Schritt soll zudem mindestens ein Zwischenschritt benannt werden.

## **Aufgabe 3** (Bayessche Parameterschätzung)

Es seien unabhängige und normalverteilte Zufallsvariablen mit $\mu = 3$ und $\sigma = 1$ gegeben. Der Parameter $\mu$ soll als unbekannt betrachtet werden und über ein Schätzverfahren approximiert werden.

Dazu ist $\mu$ mit einem bayesschen Verfahren über eine Stichprobe der Größe $2^i \cdot 10$, $i \in \{0, 1, 2, 3, 4, 5\}$, zu schätzen. Für die Bayesschätzung bei Normalverteilung gilt

$$\mu_n := \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \cdot \widehat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \cdot \mu_0.$$

a) Beschreibe zunächst die Bedeutung der folgenden Parameter:

- $\widehat{\mu}_n$,
- $\mu_0$,
- $\sigma_n^2$, definiert durch
$$\sigma_n^2 := \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2},$$
- $\sigma^2$.

Erkläre das Verhalten von $\widehat{\mu}_n$ und $\sigma_n^2$ für größer werdendes $n$ und gehe dabei auch auf $n = 0$ und kleine $n$ ein.

b) Lese die zur Verfügung gestellte Datei `samples.csv` ein.

c) Wähle für jedes $i$ die ersten $2^i \cdot 10$ Einträge aus dem Stichprobenvektor und berechne den Bayes-Schätzer für $\mu$. Für $\mu_0$ sei der Wert $-10$ angesetzt und für $\sigma_0^2$ der Wert 1.

d) Plotte für jedes $i$ jeweils in einer gemeinsamen Grafik die zu approximierende Funktion ($\mathcal{N}(3, 1)$) und deren Bayessche Schätzung. Dafür ist auch der Parameter $\sigma_n^2$ zu berücksichtigen.

e) Berechne für jedes $i$ die $L^2$-Distanz gegenüber der ursprünglichen Verteilung ($\mathcal{N}(3, 1)$) und gib sie an.

f) Beschreibe in eigenen Worten, wie sich die Bayessche Schätzung im Vergleich zur Maximum-Likelihood-Schätzung verhält.

### Hinweise

- Die Bayessche Parameterschätzung wird in Duda, S. 115–117, gut dargestellt. Die Lektüre ist zum besseren Verständnis zu empfehlen.
- Die Variation von $\mu_0$ und die Beobachtung der Auswirkungen auf die Plots kann für das Verständnis ebenfalls hilfreich sein.