GRUNDLAGEN DES MASCHINELLEN LERNENS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Tobias Lang (t.lang@uni-tuebingen.de)
Mathias Schickel (msch@fa.uni-tuebingen.de)

Andreas Schilling
Sommersemester 2020

Übungsblatt 1

30. April 2020

Organisatorisches und Formales

- Die Abgabe der Übungsblätter erfolgt in Gruppen aus **vier** Studenten auf Ilias.
- Es gibt **genau ein** Gruppenmitglied die Lösung in Form eines Archivs ab, am besten im `zip`-Format.
- Das Archiv und darin das PDF(!) mit der Textabgabe muss der Namenskonvention `nachname1-nachname2-nachname3-uebungsblattX` entsprechen. Sonderzeichen dürfen dabei nicht verwendet werden und die Namen sollen alphabetisch sortiert sein, z. B. `maier-mueller-schmidt-uebung1.zip`. Bezüglich der Abgaben der Programmieraufgaben ist zudem der letztgenannte Punkt der hier vorliegenden Anweisungen zu beachten.

  Abgaben, die diesen Konventionen nicht entsprechen, werden **nicht bewertet** und gelten als nicht erfolgt.
- Auf gute und prägnante Formulierung sowie Rechtschreibung und Zeichensetzung ist zu achten. Ausschweifende Antworten sowie mangelnde sprachliche Sorgfalt können mit Punktabzug geahndet werden.
- Zur Bearbeitung der Programmieraufgaben der Übungsblätter wird *Matlab* mit der *Statistics Toolbox* vorausgesetzt.
- 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 (Grundbegriffe und Fragen zur Vorlesung)

Beantworte die Fragen c)–f) in jeweils nicht mehr als drei Sätzen und die Übrigen so prägnant wie möglich.

a) Beantworte folgende Fragen zu den grundlegenden Definitionen der Stochastik.

   (i) Was sind *diskrete* bzw. *kontinuierliche Zufallsvariablen* X?

   (ii) Was sind in diesem Zusammenhang *Wahrscheinlichkeitsfunktion* ϱ und *Wahrscheinlichkeitsdichte* f und welcher Zusammenhang besteht jeweils zu den *Wahrscheinlichkeiten* $\mathbb{P}(\{X \in A\})$ für *Ereignisse* $\{X \in A\}$ (A eine Teilmenge des Wertebereichs S von X)?

   Wie bestimmt man also insbesondere mithilfe der Wahrscheinlichkeitsfunktion ϱ die Wahrscheinlichkeit $\mathbb{P}(\{X_i \in A_i\})$ des Ereignisses $\{X_1 \in \{a_1, \dots, a_n\}\}$ und wie mit der Wahrscheinlichkeitsdichte f die des Ereignisses $\{X_2 \in \mathbf{(}a, b\mathbf{)}\}$, $a, b \in \mathbb{R}$ mit $a \leqslant b$?

   Weise dabei darauf hin, wann es sich hier jeweils um diskrete oder kontinuierliche Zufallsvariablen handelt.

   (iii) Liefert die Multiplikation zweier Wahrscheinlichkeitsdichten wieder eine Wahrscheinlichkeitsdichte? Begründe die Antwort (und liefere gegebenenfalls ein Gegenbeispiel)!

   (iv) Wie ist der Erwartungswert $\mathbb{E}(X)$ einer Zufallsvariablen X definiert und was bedeutet er (jeweils diskret und kontinuierlich; im kontinuierlichen Fall besitze die Verteilung von X die Wahrscheinlichkeitsdichte f)?

   (v) Wie sind Varianz $\mathrm{Var}(X)$ und Standardabweichung σ einer Zufallsvariable X definiert und was bedeuten sie (diskret und kontinuierlich)? Als der Erwartungswert welcher Zufallsvariable Y lässt sich die Varianz $\mathrm{Var}(X)$ schreiben?

   (vi) Wie ist die Kovarianz $\mathrm{Cov}(X, Y)$ zweier Zufallsvariablen X und Y definiert? Was bedeutet sie? Beschreibe bitte insbesondere die Fälle $\mathrm{Cov}(X, Y) > 0, = 0$ und < 0.

b) Was ist der Unterschied zwischen *Regression* und *Klassifikation*?

c) Angenommen man möchte anhand des Abstandes von Mittelfingerspitze und Ellbogen die Größe eines Menschen bestimmen. Bietet sich hier ein Klassifikations- oder Regressionsverfahren an?

d) Wenn man analog das Geschlecht eines Menschen bestimmen möchte, bietet sich dann ein Klassifikations- oder Regressionsverfahren an?

e) Was ist der Unterschied zwischen *Supervised* und *Unsupervised Learning*?[1]

[1] Siehe dazu auch die am besten bewertete Anwort auf folgender Seite: `http://stackoverflow.com/questions/1832076/what-is-the-difference-between-supervised-learning-and-unsupervised-learning`.

Aufgabe 2 (Univariate Normalverteilung)

Gegeben sei eine reellwertige und standardnormalverteilte Zufallsvariable X (also eine normalverteilte Zufallsvariable mit Parametern $\mathbb{E}(X) =: \mu := 0$ und $\sqrt{\mathrm{Var}(X)} =: \sigma := 1$). Die Wahrscheinlichkeitsdichte f für eine normalverteilte Zufallsvariable X mit Parametern μ und σ ist dabei gegeben durch

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

a) Plotte die Dichte der Standardnormalverteilung mit den oben genannten Parametern im Intervall $(-8, 8)$.

b) Verfahre wie in a), setze allerdings jeweils einmal $\mu := -2$ und $\sigma := 2$. (Es sind in diesem Aufgabenteil also *drei* Plots zu erstellen.) Erkläre die Änderung der Plots in zwei bis drei Sätzen und erläutere dabei die Bedeutung der beiden Parameter μ und σ für den Graphen der Wahrscheinlichkeitsdichte f der Normalverteilung.

c) Mit welcher (gerundeten) Wahrscheinlichkeit nimmt X Werte im Intervall $I := (\mu - \sigma, \mu + \sigma)$ an?

d) Ist es für eine Wahrscheinlichkeitsdichte f möglich, dass der Wert $f(x)$ an einer Stelle[2] x des Definitionsbereiches von f größer als 1 ist? Begründe die Antwort!

Hinweise

- In Matlab kann die Funktion `normpdf` (`http://de.mathworks.com/help/stats/normpdf.html`) verwendet werden.
- Zum Plotten von Funktionen bietet sich `fplot`, `http://de.mathworks.com/help/matlab/ref/fplot.html`, an.

[2] Bzw. innerhalb einer Teilmenge M mit Lebesguemaß $\lambda(M) > 0$ (für die Kenner der Maßtheorie).

Aufgabe 3 (Multivariate Normalverteilung)

Die *multivariate Normalverteilung* ist die gemeinsame Verteilung mehrerer Zufallsvariablen X_i, $i \in I$. Zusätzlich zu den Parametern $\mu_i := \mathbb{E}(X_i)$ und $\sigma_i := \sqrt{\mathrm{Var}(X_i)}$ spielen bei der multivariaten Normalverteilung die Kovarianzen $\mathrm{Cov}(X_i, X_j)$, $i, j \in I$, eine Rolle. Im Folgenden soll der Einfluss der Kovarianz auf den Graphen der Wahrscheinlichkeitsdichte der multivariaten Normalverteilung zweier Zufallsvariablen $X := X_1$ und $Y := X_2$ anschaulich untersucht werden. Für diese Aufgabe soll das Matlabskript `myMvmPdf.m` verwendet werden, das zusammen mit diesem Übungsblatt zur Verfügung gestellt wird. Die Teile des Skripts, die bearbeitet werden sollen, sind mit einem `TODO` gekennzeichnet.

a) Verändere die Einträge des Vektors μ und verwende dabei Werte aus dem Intervall $(-2, 2)$. Welche Konsequenzen hat die Änderung? Eine Erklärung und ein Beispielplot sind erforderlich!

b) Der Vektor μ soll nun wieder auf 0 gesetzt werden. Setze die Variable `covarianceX1X2` jeweils einmal auf 0.7, 0.99, −0.7 und −0.99. Was kann man an den (vier) Plots ablesen (Erklärung und jeweils ein Beispielplot)?

c) Die Kovarianz $\mathrm{Cov}(X, Y)$ der Zufallsvariablen X und Y unterscheidet sich vom *Korrelationskoeffizienten* $\mathrm{Korr}(X, Y)$. Schlage hierzu `https://de.wikipedia.org/wiki/Korrelationskoeffizient#Bildliche_Darstellung_und_Interpretation` nach und beschreibe grob den genannten Unterschied.

Hinweis Die Kovarianzmatrix ist stets quadratisch, symmetrisch und positiv semidefinit. Diese Eigenschaften limitieren die Wahl der Einträge der Matrix im Skript. (Vgl. auch `http://de.wikipedia.org/wiki/Definitheit#Definitheit_von_Matrizen`.)

Aufgabe 4 (Modellierung von Zufallsvariablen)

Häufig kommt es vor, dass für ein Zufallsexperiment nicht bekannt ist, welche Wahrscheinlichkeitsfunktion ϱ die zugehörige Zufallsvariable X charakterisiert. Betrachte als Beispiel ein Zufallsexperiment mit einem gezinkten zehnseitigen Würfel. Es soll statistisch eine Wahrscheinlichkeitsfunktion ϱ ermittelt werden.

Im File `wuerfel.csv` sind die Ergebnisse von 100 Würfen mit dem gezinkten Würfel notiert. Außerdem werden für diese Aufgabe Funktionsrümpfe mitgeliefert (im Unterordner `A4` des Ordners `data`), die für die Aufgabe verwendet werden müssen. Gehe nun folgendermaßen vor:

a) Lese das File in Matlab ein. (Der Eintrag in jeder Zeile steht für die Augenzahl des entsprechenden Wurfs.)

b) Plotte das Histogramm für die Daten.

c) Plotte die diskrete Wahrscheinlichkeitsfunktion ϱ. (Überlegt Euch dazu, wie man eine solche aus dem Histogramm des vorangegangenen Aufgabenteils erhält.)

d) Erstelle und plotte die diskrete Verteilungsfunktion F für die Wahrscheinlichkeitsfunktion ϱ. Der Wert der Verteilungsfunktion F an der Stelle x, $x \in \{1, \dots, 10\}$,

$$F(x) \coloneqq \mathbb{P}(X \leqslant x),$$

gibt an, mit welcher Wahrscheinlichkeit die Zufallsvariable X einen Wert zwischen 1 und x annimmt.

e) Mittels der Verteilungsfunktion F können Zufallszahlen erzeugt werden, deren Verteilung der des Würfels entspricht. Gehe dazu folgendermaßen vor:

- Erzeuge eine uniforme Zufallszahl im Intervall $(0, 1)$.
- Suche beginnend beim letzten Eintrag von F den ersten, der kleiner als die erzeugte Zufallszahl ist.
- Das Ergebnis ist der nächsthöhere Index.

Auf diese Weise wird eine uniform verteilte Zufallszahl im Intervall $(0, 1)$ modelliert und als Argument für die inverse Verteilungsfunktion verwendet.

f) Angenommen man ist nicht zufrieden mit den Ergebnissen des obigen Zufallszahlengenerators. Wie könnte man präzisere Ergebnisse erhalten?

Hinweise

- Zum Einlesen von `.csv`-Files in Matlab bietet sich die Funktion `csvread` an.
- Eine uniforme Zufallszahl kann in Matlab mittels des Befehls `unifrnd(min, max)` erstellt werden.
- Für das Erstellen eines Histogramms bietet sich in Matlab die Funktion `hist` an. (Folgt kein Semikolon am Ende der Zeile, erzeugt sie einen Plot.)

- Das Plotten der Wahrscheinlichkeitsfunktion ϱ und der Verteilungsfunktion F lässt sich zum Beispiel mittels der Matlabfunktion `bar(…)` umsetzen.
- Hilfreich zum Verständnis kann auch folgender Wikipediaartikel sein: `http://de.wikipedia.org/wiki/Inversionsmethode`.