Prof. Dr. Andreas Schilling
Tobias Lang (`t.lang@uni-tuebingen.de`)
Mathias Schickel (`msch@fa.uni-tuebingen.de`)

Tübingen, den 14.04.2016

Übungen zur Vorlesung Maschinelles Lernen

Übungsblatt 1

Ausgabe: 15.04.2016; Abgabe: bis 21.04.2016, 23:59 Uhr.

Organisatorisches

- Die Abgabe der Übungsblätter erfolgt (spätestens ab Übungsblatt 2) in Gruppen aus drei Studenten auf Ilias.
- Dazu gibt bitte immer nur genau ein Gruppenmitglied die Lösung in Form eines Archivs in Ilias ab.
- Das Archiv sollte folgender Namenskonvention entsprechen (keine Großbuchstaben) und aus der Lösungsdatei sollte zudem klar das Team hervorgehen: nachname1nachname2nachname3uebung$X$.archivformat, wobei X die Nummer des Übungsblattes bezeichne. Sonderzeichen sollten dabei nicht verwendet werden. Beispiel: schmidtmuellermaieruebung1.tar.gz.
- Zur Bearbeitung der Programmieraufgaben der Übungsblätter wird *Matlab* mit der *Statistics Toolbox* vorausgesetzt.

Aufgabe 1 (Fragen zur Vorlesung) (16 Punkte)

Beantworte bitte die folgenden Fragen in jeweils nicht mehr als drei Sätzen:

a) Was ist der Unterschied zwischen Regression und Klassifikation? (1 Punkt)

b) Angenommen man möchte anhand des Abstandes von Mittelfingerspitze und Ellbogen die Größe eines Menschen bestimmen. Bietet sich hier ein Klassifikations- oder Regressionsverfahren an? (1 Punkt)

c) Angenommen man möchte analog das Geschlecht eines Menschen bestimmen. Bietet sich hier ein Klassifikations- oder Regressionsverfahren an? (1 Punkt)

d) Was ist der Unterschied zwischen *Supervised* und *Unsupervised Learning*?[1] (1 Punkt)

e) Wiederhole und nenne die grundlegenden Definitionen der Wahrscheinlichkeitsrechnung:

  (i) Was ist eine *diskrete* und was eine *kontinuierliche Zufallsvariable* X? Was ist ein *Ereignis*? (2 Punkte)

  (ii) Worin unterscheidet sich der diskrete vom kontinuierlichen Fall? 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 $\{X \in A\}$ („X fällt in A“), A eine Teilmenge des Wertebereichs S von X? (2 Punkte)

  (iii) Wie ist der Erwartungswert $\mathbb{E}(X)$ einer Zufallsvariable X definiert und was bedeutet er (jeweils diskret und kontinuierlich, wobei im kontinuierlichen Fall vorausgesetzt werden darf, dass die Verteilung von X eine Wahrscheinlichkeitsdichte f besitzt)? (2 Punkte)

  (iv) Wie sind Varianz $\mathrm{Var}(X)$ und Standardabweichung σ einer Zufallsvariable X definiert und was bedeuten sie (diskret und kontinuierlich wie oben)? Lässt sich die Varianz $\mathrm{Var}(X)$ als Erwartungswert $\mathbb{E}(Y)$ einer geeigneten Zufallsvariable Y schreiben? Falls ja: Was ist Y? (3 Punkte)

  (v) 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. (2 Punkte)

f) Liefert die Multiplikation zweier Wahrscheinlichkeitsdichten wieder eine solche? Begründung! (2 Punkte)

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

Aufgabe 2 (Univariate Normalverteilung) (12 Punkte)

Gegeben sei eine reellwertige und standardnormalverteilte Zufallsvariable X (also eine normalverteilte Zufallsvariable mit Parametern $\mathbb{E}(X) =: \mu := 0$ und $\sqrt{\operatorname{Var}(X)} =: \sigma := 1$). 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 Standardnormalverteilung mit den oben genannten Parametern im Intervall $[-8, 8]$. (2 Punkte)

b) Verfahre wie in a), setze allerdings jeweils einmal $\mu := -2$ und $\sigma := 2$. Begründe die Änderung des Plots in zwei bis drei Sätzen. (2 Punkte)

c) Was bedeutet der Wert des Parameters $\mu = \mathbb{E}(X)$ für die Normalverteilung? (1 Punkt)

d) Welche qualitative Bedeutung hat der Parameter $\sigma = \sqrt{\operatorname{Var}(X)}$ für die Normalverteilung? (1 Punkt)

e) Wie bestimmt man mithilfe der Wahrscheinlichkeitsdichte f die Wahrscheinlichkeit des Ereignisses $\{X \in [a, b]\}$, $\mathbb{P}(X \in [a, b])$, $a, b \in \mathbb{R}$ mit $a \leqslant b$? (2 Punkte)

f) Mit welcher Wahrscheinlichkeit nimmt X Werte an in den Intervallen
   (i) $I_1 := [\mu - \sigma, \mu + \sigma]$,
   (ii) $I_2 := [\mu - 2\sigma, \mu + 2\sigma]$,
   (iii) $I_3 := [\mu - 3\sigma, \mu + 3\sigma]$? (zusammen 2 Punkte)

g) Ist es für eine Wahrscheinlichkeitsdichte f möglich, dass ihr Wert $f(x)$ an einer Stelle x ihres Definitionsbereiches größer als 1 ist? Begründe die Antwort! (2 Punkte)

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.

Aufgabe 3 (Multivariate Normalverteilung) (8 Punkte)

In der Vorlesung wurde die *multivariate Normalverteilung* vorgestellt. Eine multivariate Normalverteilung kann, im Gegensatz zur *univariaten Normalverteilung*, mehrere Variablen X_i, $i \in I$, und deren Zusammenhang berücksichtigen. Während die Parameter $\mu_i \coloneqq \mathbb{E}(X_i)$ und $\sigma_i \coloneqq \sqrt{\mathrm{Var}(X_i)}$ bereits in der vorherigen Aufgabe behandelt wurden, spielen bei der multivariaten Normalverteilung zudem die Kovarianzen $\mathrm{Cov}(X_i, X_j)$, $i, j \in I$, eine Rolle. Im Folgenden soll anschaulich der Einfluss der Kovarianz auf den Graphen der Wahrscheinlichkeitsdichte der multivariaten Normalverteilung zweier Zufallsvariablen $X \coloneqq X_1$ und $Y \coloneqq X_2$ 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) Verändert die Einträge des Vektors μ und verwendet dabei Werte aus dem Intervall $[-2, 2]$. Welche Konsequenzen hat die Änderung? Erklärung und ein Beispielplot! (3 Punkte)

b) Der Vektor μ soll nun wieder auf 0 gesetzt werden. Setzt die Variable `covarianceX1X2` dann einmal auf 0.7, 0.99, −0.7 und −0.99. Was kann man an den Plots ablesen? Erklärung und jeweils ein Beispielplot. (3 Punkte)

c) Die Kovarianz $\mathrm{Cov}(X, Y)$ der Zufallsvariablen X und Y unterscheidet sich von deren *Korrelationskoeffizient*. Schlagt hierzu `http://de.wikipedia.org/wiki/Korrelationskoeffizient#Bildliche_Darstellung_und_Interpretation` nach und beschreibt grob den genannten Unterschied. (2 Punkte)

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

Aufgabe 4 (Generierung von eigenen Zufallsvariablen) (8 Punkte)

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 10-seitigen Würfel. Nun soll statistisch eine Wahrscheinlichkeitsfunktion ϱ ermittelt werden. Im File `wuerfel.csv` seien die Ergebnisse von 100 Würfen mit dem gezinkten Würfel notiert. Geht nun folgendermaßen vor:

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

b) Plottet das Histogramm für die Daten. (1 Punkt)

c) Plottet die diskrete Wahrscheinlichkeitsfunktion ϱ. (Überlegt Euch dazu, wie man eine solche aus dem Histogramm des vorangegangenen Aufgabenteils erhält.) (2 Punkte)

d) Erstellt und plottet 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. (2 Punkte)

e) Mittels der Verteilungsfunktion F könnt ihr nun Zufallszahlen erzeugen, deren Verteilung der des Würfels entspricht. Geht dazu folgendermaßen vor:

- Erzeugt eine uniforme Zufallszahl im Intervall $[0, 1]$.
- Sucht beginnend beim letzten Eintrag von F den ersten, der kleiner als die erzeugte Zufallszahl ist.
- Das Ergebnis ist der nächst höhere Index.

Mathematisch gesehen erstellt man so eine uniform verteilte Zufallszahl im Intervall $[0, 1]$ und verwendet diese als Argument für die inverse Verteilungsfunktion. (2 Punkte)

f) Angenommen Ihr seid nicht zufrieden mit den Ergebnissen des obigen Zufallszahlengenerators. Was könntet Ihr tun, um präzisere Ergebnisse zu erhalten? (1 Punkt)

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`.