GRUNDLAGEN DES MASCHINELLEN LERNENS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

Andreas Schilling
Sommersemester 2020

Blatt 2

Ausgabe: 15.05.2020; Abgabe: bis 29.05.2020, 09:00 Uhr.

I don't believe in the idea that there are a few peculiar people capable of understanding math, and the rest of the world is normal.
RICHARD P. FEYNMAN

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 darin das PDF(!) mit der Textabgabe muss 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 (Bedingte Wahrscheinlichkeiten und der Satz von Bayes)

Wenn nach Definitionen oder Sätzen gefragt wird, sollte stets die Formel angegeben werden.

a) Wie lässt sich für zwei *stochastisch unabhängige* Ereignisse A und B die Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ („die Wahrscheinlichkeit von A und B“) bestimmen?

b) Wie ist die *bedingte Wahrscheinlichkeit* $\mathbb{P}(A \mid B)$ („die Wahrscheinlichkeit von A unter der Bedingung B“ oder „gegeben B“) in den Termen von $\mathbb{P}(A \cap B)$ und $\mathbb{P}(B)$ definiert und was bedeutet sie? Welcher Wahrscheinlichkeit gleicht $\mathbb{P}(A \mid B)$, wenn A und B stochastisch unabhängig sind?

c) Wie lautet der *Satz von Bayes* und wie nennt man die in der Formel vorkommenden einzelnen Terme? Weise den Satz von Bayes nach und verwende dazu die Definition aus der vorangegangenen Aufgabenstellung.

d) Wie lautet der *Satz von der totalen Wahrscheinlichkeit* für ein Ereignis B und paarweise disjunkte Ereignisse A_i mit $\dot{\bigcup}_i A_i = \Omega$ (für eine Grundmenge Ω)?

e) Die (absolute) Wahrscheinlichkeit(sdichte) $P(x)$ gegebener Daten (auch bezeichnet als *evidence*) ist hypothesenunabhängig. Warum kann der Nenner im Satz von Bayes deswegen beim Vergleich verschiedener Hypothesen vernachlässigt werden? Wann ist er wichtig? (Beachte dabei Folie 40 in Foliensatz 2.)

f) Warum *summieren* sich auf Folie 38 (Foliensatz 2) die Funktionswerte *punktweise* (d. h. für jedes *einzelne* x) genau zu 1 und auf Folie 36 nicht? Sind die entsprechenden *Integrale* über den Werte*bereich* von x jeweils 1? (Überlege dazu, um die Graphen welcher Funktionen es sich jeweils handelt.)

Aufgabe 2 (Entscheidungstheorie)

Wenn nach Definitionen gefragt wird, sollte stets die Formel angegeben werden. Außerdem sollen die Fragen möglichst prägnant beantwortet werden.

a) Was ist das allgemeine Ziel der automatisierten Entscheidungsfindung und auf welcher Basis sollen automatische Entscheidungen getroffen werden?

b) Wann liegt ein *Fehler* bei einer solchen Entscheidungsfindung vor? Sollten alle Fehlentscheidungen gleichermaßen behandelt werden?

c) Wie sollte zwischen zwei Hypothesen $\omega_1, \dots, \omega_n$ entschieden werden, wenn ausschließlich die Auftrittswahrscheinlichkeiten $\mathbb{P}(\omega_j)$, $i = 1, \dots, n$, bekannt sind? Begründe die Antwort.

d) Wie entscheidet man sinnvollerweise in vorliegenden Datenpunkten x, wenn die bedingten Auftrittswahrscheinlichkeiten $\mathbb{P}(\omega_j \mid x)$, $i = 1, \dots, n$, bekannt sind?

   Nenne dabei die bayessche Entscheidungsregel und erläutere sie. Was wird durch sie erreicht? (Siehe dazu Foliensatz 2, Folie 39 oder Duda S. 42.)

e) Was besagt der Wert der *Verlustfunktion* $\lambda(\omega_i \mid \omega_j)$ für ω_i und ω_j? (Siehe Foliensatz 2, S. 41.) Wozu dient die Verlustfunktion? (Beachte dazu auch Aufgabenteil b.)

f) Wie ist das *(bedingte) Risiko* (conditional risk) $\mathcal{R}(\omega_i \mid x)$ der Entscheidung für ω_i im Datenpunkt x definiert (Formel) und was bedeutet es? (Siehe Foliensatz 2, S. 41 und auch Duda S. 43 /44.)

g) Wie ist das *Gesamtrisiko* $\mathfrak{R}(\omega)$ einer Entscheidungsfunktion $\omega : X \to \Omega$ definiert? Was besagt das Gesamtrisiko für eine Entscheidungsfunktion ω? Wie lautet die optimale Entscheidungsregel und wie ist das *bayessche Risiko* definiert? (Siehe Foliensatz 2, Folien 42 und 43 bzw. Duda S. 44.)

   Liefert die optimale Entscheidungsregel *in jeder konkreten Situation* das beste Ergebnis? Ist es möglich, eine Entscheidungsfunktion zu finden, die mit einem geringeren Gesamtrisiko als dem bayesschen Risiko verbunden ist?

Hinweis Für das Verständnis ist die Lektüre des Duda, PDF-S. 39–45 in der digitalen Version bzw. S. 20–27 im gedruckten Buch, zu empfehlen.

Aufgabe 3 (Bayesscher Fischklassifikator)

Ein Fischer hat dir neulich eine Liste zukommen lassen (`fische.csv`), in der die Längen der Fische seines letzten Fanges aufgelistet sind. Wir treffen als Grundlage für die folgende Untersuchen diese Annahmen:

- Es gibt gleich viele Lachse und Barsche. Zudem gibt es auch *nur* Lachse und Barsche.
- Die Länge der Barsche (gemessen in Metern) ist normalverteilt mit $\mu_1 := 1$ und $\sigma_1 := 0.2$ und die der Lachse mit $\mu_2 := 1.6$ und $\sigma_2 := 0.3$.

Nun sollen die Fische des Fangs des Fischers klassifiziert werden.

a) Lese die Längen der Fische aus dem `csv`-File in Matlab ein. Die Längen stehen dabei in der ersten Spalte.

b) Plotte zunächst (zur eigenen Orientierung) ein Histogramm für die Häufigkeit der Fischlängen des Fangs. Wie verlässlich (subjektiv) ist die oben getroffene Annahme zu den Mittelwerten in der Aufgabenstellung auf der Basis des Histogramms?

c) Wende nun für die Länge x den *Satz von Bayes* an, um zu entscheiden, ob es sich bei einem Fisch dieser Länge um einen Barsch (ω_1) oder einen Lachs (ω_2) handelt.

   (i) Berechne die *likelihood* $p(x \mid \omega_j)$ eines Barsches (ω_1) oder Laches (ω_2) für die Länge x (d. h. den Wert der bedingten Wahrscheinlichkeitsdichte für die Länge x, gegeben, dass ein Barsch (ω_1) oder ein Lachs (ω_2) vorliegt).[1]

   (ii) Bestimme die *evidence*.

   (iii) Auf der Basis dieser Werte kann anhand des Satzes von Bayes die *a-Posteriori Wahrscheinlichkeit* bestimmt werden, dass es sich bei einem Fisch der gegebenen Länge x um einen Barsch (ω_1) oder Lachs (ω_2) handelt.

d) Plotte die bedingten Funktionen $\mathbb{P}(\omega_j \mid x)$, $j = 1, 2$, (Klasse *Barsch* und Klasse *Lachs*). Erkläre (kurz), wie man anhand des Plots die Klassifikation durchführt.

e) Entscheide anhand der *Bayes'schen Entscheidungsregel* (siehe dazu auch Aufgabe 1), um welche Art Fisch es sich bei den jeweiligen Exemplaren aus dem Datensatz vermutlich handelt.

f) Wie wahrscheinlich ist das Vorliegen von Barsch oder Lachs in der Stichprobe auf der Basis des Klassifikationsergebnisses? Gib die Wahrscheinlichkeiten als Kommentar im Code an und beurteile das Ergebnis.

g) Erläutere kurz die Abhängigkeit der Klassifikation von den getroffenen Annahmen.

[1] Die Terminologie ist zunächst etwas verwirrend und bezieht sich auf S. 41 des Duda (PDF-Version, in der Druckversion S. 22): Die bedingte Wahrscheinlichkeitsdichte $p(x \mid \omega_j)$ heißt dort *likelihood of* ω_j *given that the feature value* x *has been measured*. Gleichwohl handelt es sich in der Terminologie der *Stochastik* um die *bedingte Wahrscheinlichkeitsdichte* des *features* x *gegeben* ω_j (also hier der Länge x, wenn ω_j bekannt ist).

Hinweise

- In Matlab kann die Funktion `csvread('dateiname')` zum Einlesen von `csv`-Dateien verwendet werden.
- Zum Plotten von Datenpunkten kann die Funktion `scatter()` verwendet werden.
- Als Plotintervall bietet sich `[0.4 2.0]` an.
- Die Matlabfunktion `hist(x, numBins)` plottet ein Histogramm der Daten im Vektor `x` und liefert dessen Daten (Säulenhöhen) eingeteilt in `numBins` Abschnitte zurück.

Aufgabe 4 (Verlustfunktion und Risiko)

Angenommen, eine Regulierungsbehörde entscheidet, den Lachs auf die Liste der schützenswerten Arten zu setzen. Beim Fischfang müssen daher gefangene Lachse zurück ins Meer geworfen werden. Für die Fischereien ist es deswegen wichtig, dass Lachse so selten wie möglich falsch klassifiziert werden. Zu diesem Zweck muss der Klassifikator aus der letzten Aufgabe entsprechend überarbeitet werden. Hierzu sei eine *Loss-Funktion* λ gegeben durch

$$\begin{aligned}\lambda(\text{Barsch} \mid \text{Barsch}) &:= 0,\\ \lambda(\text{Barsch} \mid \text{Lachs}) &:= 1.2,\\ \lambda(\text{Lachs} \mid \text{Barsch}) &:= 0.5,\\ \lambda(\text{Lachs} \mid \text{Lachs}) &:= 0.\end{aligned}$$

Der Wert $\lambda(x \mid y)$ gibt die *Kosten* der Fehlklassifikation an, wenn *tatsächlich* y vorliegt, aber $x \neq y$ *klassifiziert* wird. Mithilfe dieser *Loss-Funktion* sollen beim obigen Klassifikator Lachs-Fehlklassifikationen vermieden werden. Gehe dazu folgendermaßen vor:

a) Berechne nach der Feststellung der a-Posteriori Wahrscheinlichkeit (in der letzten Aufgabe) das Risiko (*conditional risk*) für die *irrtümliche* Entscheidung, dass ein Fisch einer bestimmen Länge als Barsch betrachtet wird, sowie das analoge Risiko für Lachs.

b) Plotte das *conditional risk* jeweils für Barsch und Lachs und überlege, wie anhand des Plots entschieden werden kann, ob Fische gegebener Längen als Lachs oder Barsch zu klassifizieren sind.

c) Entscheide auf der Basis der im vorangegangen Aufgabenteil entwickelten Methode, um welche Klasse Fisch es sich bei den jeweiligen Exemplaren aus dem Datensatz handelt.

d) Ändere die *Loss Funktion* auf eine beliebige Art ab. Beschreibe, wie sich das Klassifikationsverhalten ändert.

Aufgabe 5 (Bayesscher multivariater Klassifikator)

Ein klassischer multivariater Datensatz beinhaltet Kenngrößen dreier verschiedener Schwertlilienarten.[2] Mittels eines bayesschen Klassifikators sollen die Daten den drei Arten zugeordnet werden, wobei zunächst von der *stochastischen Unabhängigkeit* der Kenngrößen der einzelnen Arten ausgegangen werden soll. Dazu stehen Trainings- und Testdaten zur Verfügung, und zwar einmal in den Files `trainingSetosa.csv`, `trainingVersicolor.csv` und `trainingVirginica.csv` sowie in `testSetosa.csv`, `testVersicolor.csv` und `testVirginica.csv`. Ein solcher Klassifikator soll anhand der folgenden Schritte erstellt werden.

a) Zunächst sollen die Trainings- und Testdaten für jede Schwertlilienart in Matlab eingelesen werden. Jede Spalte entspricht jeweils einer Messreihe (d. h. in den Zeilen finden sich die Messwerte) für eine der vier Kenngrößen[3]

   1. *sepale Länge* (Länge des Kelchblattes),
   2. *sepale Breite* (Breite des Kelchblattes),
   3. *petale Länge* (Länge des Kronblattes),
   4. *petale Breite* (Breite des Kronblattes).

   Für die Testdaten ist zwar bekannt, zu welcher Schwertlilienart sie jeweils gehören. Dieses Wissen darf beim Testen des Klassifikators aber selbstverständlich *ebensowenig* verwendet werden wie die *Testdaten* zum *Trainieren* des Klassifikators berücksichtigt werden dürfen.

b) Die Daten zu jeder Kenngröße jeder Schwertlilienart aus den *Trainings*daten sollen in einem Histogramm veranschaulicht werden. Höchstens sechs Sätze sollen zudem Auffälligkeiten in den Daten beschreiben.

c) Für jede Kenngröße jeder Schwertlilienart sind anschließend sowohl der Mittelwert als auch die Varianz für die *Trainings*daten zu bestimmen.

d) Bestimme die jeweiligen *likelihoods* $p(x \mid \omega)$ für die Schwertlilienarten. Vorausgesetzt werden soll dazu, dass die einzelnen Kenndaten jeweils normalverteilt sind mit den in der letzten Teilaufgabe ermittelten jeweiligen Mittelwerten bzw. Standardabweichungen. Danach kann der *Satz von Bayes* zur Klassifikation verwendet werden.

e) Für jede Schwertlilienart sind folgende Maßzahlen zu bestimmen:

   - Die Anzahl der Datenpunkte, die *korrekt* als *zur Art gehörig* klassifiziert wurden (*true positive*).
   - Die Anzahl der Datenpunkte, die *korrekt* als *nicht zur Art gehörig* klassifiziert wurden (*true negative*).

---

[2] Vgl. dazu auch `http://en.wikipedia.org/wiki/Iris_flower_data_set` und `http://archive.ics.uci.edu/ml/datasets/Iris`.

[3] Zum näheren Verständnis der einzelnen Kenngrößen siehe etwa `http://de.wikipedia.org/wiki/Kelchblatt` und `http://de.wikipedia.org/wiki/Kronblatt`.

- Die Anzahl der Datenpunkte, die *fälschlicherweise* als *zur Art gehörig* klassifiziert wurden (*false positive*).
- Die Anzahl der Datenpunkte, die *fälschlicherweise* als *nicht zur Art gehörig* klassifiziert wurden (*false negative*).

f) Folgende Kenngrößen sind in Grafiken abzutragen und in maximal sechs Sätzen zu beurteilen:
   - Petrale Breite gegen sepale Breite,
   - petale Länge gegen sepale Länge,
   - petale Länge gegen petale Breite,
   - sepale Länge gegen sepale Breite.

g) Bislang ist für die einzelnen Kenndaten jeder Schwertlilienart stochastische Unabhängigkeit vorausgesetzt worden. Im Gegensatz dazu soll nun für die likelihood $p(x \mid \omega)$ der Kenndatensätze jeder einzelnen Schwertlilienart multivariate Normalverteilung angenommen werden. Auf dieser Basis soll die Klassifikation erneut durchgeführt und die Änderung des Ergebnisses beschrieben werden. Wie sind die Änderungen zu erklären? Ein Beispielplot zur Veranschaulichung ist erwünscht.

Hinweise zum Implementierung in Matlab

a) Die spaltenweise Berechnung des Mittelwertes und der Varianz ist in Matlab mittels der Funktionen `mean(<Matrix>)` und `var(<Matrix>)` möglich. Das Ergebnis ist ein Vektor mit den Mittelwerten bzw. Varianzen jeder Kenngröße in der Matrix.

b) Die Funktion `normpdf( X, mu, sigma )` kann auf Matrizen `X` mit mehr als einer Spalte angewandt werden – die Spalten stehen dabei jeweils für eine Kenngröße und die Zeilen für Messwerte – und erzeugt die Werte der Wahrscheinlichkeitsdichte für die Normalverteilung mit Parametern `mu` und `sigma` an den durch die Zeilen der Matrix `X` gegebenen Messwerten. Dabei müssen `mu` und `sigma` jeweils selbst Matrizen sein, die in jeder Spalte den Mittelwert- und die Standardabweichung der zugehörigen Kenngröße enthalten. Die Zeilenzahl muss derjenigen der Matrix entsprechen. Nützlich dafür ist der Befehl `repmat` (eventuell nur bei älteren Matlab-Versionen nötig). Die Ausgabe ist eine Matrix `Y` mit den Werte der jeweiligen Dichte an den entsprechenden Stützstellen aus `X`.

c) Für die übersichtliche Gestaltung des Programmiercodes ist in Matlab an Stelle der Verwendung von Schleifen Vektor- bzw. Matrixschreibweise anzuraten – zudem läuft der Code dann performanter. Der vorangegangene Hinweis ist dafür hilfreich.

d) Die Kovarianzmatrix eines `n` × `d` Datensatzes `X` mit `n` Zeilen und `d` Kennzahlen lässt sich mittels des Befehls `cov(X)` berechnen.

e) Der Wert der Wahrscheinlichkeitsdichte für die multivariate Normalverteilung mit Mittelwertvektor `mu` und Kovarianzmatrix `C` an den durch die Matrix `X` gegebenen Datenpunkten kann in Matlab mit dem Befehl `mvnpdf( X, mu, C )` berechnet werden. Es ist eventuell nützlich, dazu die Matlab-Hilfe zu konsultieren.