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 2016

Übungsblatt 2

Ausgabe: 21.04.2016; Abgabe: bis 28.04.2016, 23:59 Uhr.

Organisatorisches

- Die Abgabe **muss** in Gruppen aus drei Personen erfolgen.
- Es gibt **nur genau ein** Gruppenmitglied ein Archiv mit allen Dateien ab.
- Das Archiv **muss** folgender Namenskonvention entsprechen (keine Großbuchstaben): `nachname1nachname2nachname3uebungX.archivformat`, wobei `X` die Nummer des Übungsblattes bezeichne. Beispiel: `schmidtmuellermaieruebung1.tar.gz`.
- **Aller** erstellter (und lauffähiger) Lösungscode wird abgegeben.

Aufgabe 1 (Fragen zur Vorlesung) (10 Punkte)

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

a) Wie lautet der *Satz von Bayes*? Welche Anwendungen dieses Satzes wurden in der Vorlesung vorgestellt? (1 Punkt)

b) Wie kann man den Satz von Bayes in Bezug auf eine Hypothese interpretieren, die man auf der Basis von ermittelten Daten prüfen möchte? (1 Punkt)

c) Die (absolute) Wahrscheinlichkeit, dass die gegebenen Daten vorliegen, (auch bezeichnet als *evidence*) ist unabhängig von der Hypothese. Im Satz von Bayes ist daher (Foliensatz 2, Folie 24) der Nenner bei der Hypothesenbewertung in gewisser Hinsicht qualitativ vernachlässigbar. Wann ist er dennoch wichtig? Fasst kurz zusammen, in welchen Fällen die evidence vernachlässigt werden kann und in welchen nicht. (1 Punkt)

d) Wie kann man zwischen zwei Hypothesen ω_1 und ω_2 entscheiden, wenn keine Daten vorliegen? ($\mathbb{P}(\omega_i)$, $i = 1, 2$, darf dabei als bekannt vorausgesetzt werden.) (1 Punkt)

e) 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? (Überlegt dazu genau, was die Grafiken zeigen.) (2 Punkte)

f) *Erkläre*, was die Bayes'sche Entscheidungsregel besagt. (Siehe dazu Foliensatz 2, Folie 39 oder Duda S. 42.) (2 Punkte)

g) Welches Risiko drückt das *(conditional) risk* aus? (Dieses findet sich in Foliensatz 2 auf S. 41; siehe auch Duda S. 43 / 44.) (1 Punkt)

h) Wie hängen das sogenannte *Bayes risk* (Foliensatz 2, Folie 43 bzw. Duda S. 44) und das *Gesamtrisiko* (Foliensatz 2, Folie 42) zusammen? (1 Punkt)

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 2 (Bayes'scher Fisch-Klassifikator) (14 Punkte)

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 normalverteilt mit $\mu_2 := 1.6$ und $\sigma_2 := 0.3$.

Nun sollen die Fische des Fangs des Fischers klassifiziert werden.

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

b) Plottet zunächst (zur eigenen Orientierung) ein Histogramm für die Häufigkeit der Fischlängen des Fangs. Für wie verlässlich haltet Ihr (subjektiv) die Annahme zu den Mittelwerten in der Aufgabenstellung auf der Basis des Histogramms?

c) Wendet nun für die Länge x den *Satz von Bayes* an, um zu entscheiden, ob es sich um einen Barsch (ω_1) oder einen Lachs (ω_2) handelt. (6 Punkte)

   (i) Berechnet für die Länge x die *likelihood* $p(x|\omega_j)$ eines Barsches (ω_1) oder Lachses (ω_2) in Bezug auf die Länge x (d. h. die bedingte Wahrscheinlichkeitsdichte der Länge x, gegeben einmal, dass ein Barsch (ω_1) vorliegt, und einmal, dass ein Lachs (ω_2) vorliegt).[1]

   (ii) Bestimmt nun die *evidence*.

   (iii) Im Anschluss könnt ihr anhand des Satzes von Bayes die *a-Posteriori Wahrscheinlichkeit* bestimmen, dass es sich bei gegebener Länge x um einen Barsch (ω_1) oder Lachs (ω_2) handelt.

d) Plottet nun die beiden bedingten Funktionen $\mathbb{P}(\omega_j|x)$, $j = 1, 2$, für die Barsche und Lachse. Erklärt (kurz), wie man anhand des Plots die Klassifikation in Barsche und Lachse durchführt. (2 Punkte)

e) Entscheidet nun anhand der *Bayes'schen Entscheidungsregel* (siehe dazu auch Aufgabe 1), um welche Art Fisch es sich bei den Fischen der einzelnen Längen vermutlich handelt. (3 Punkte)

f) Wie wahrscheinlich ist das Vorliegen eines Barsches oder Lachses in der Stichprobe, gesetzt den Fall, die Klassifikation wäre korrekt? Gebt diese Wahrscheinlichkeiten als Kommentare im Code an und beurteilt das Ergebnis. (2 Punkte)

g) Inwiefern ist das Klassifikationsverfahren abhängig von den Annahmen? (1 Punkt)

Hinweise:

- In Matlab kann die Funktion `csvread('dateiname')` zum Einlesen von `.csv`-Dateien 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.

[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 $p(x|\omega_j)$ 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).

Aufgabe 3 (Loss-Funktion, Risiko) (10 Punkte)

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. Die *Loss-Funktion* λ sei dabei gegeben als

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

Dabei gibt $\lambda(x|y)$ die *Kosten* der Fehlklassifikation an, wenn *tatsächlich* y vorliegt, aber $x \neq y$ *klassifiziert* wird. Nun sollen mithilfe dieser *Loss-Funktion* beim obigen Klassifikator Lachs-Fehlklassifikationen vermieden werden. Dazu wird folgendermaßen vorgegangen:

a) Berechnet 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 Lachse. (3 Punkte)

b) Plottet das *conditional risk* jeweils für Barsche und Lachse und überlegt, wie man anhand des Plots entscheidend kann, Fische welcher Länge man als Lachse und welche als Barsche klassifizieren sollte. (3 Punkte)

c) Entscheidet mit der Idee aus dem vorangegangen Aufgabenteil anhand des *conditional risk*, um welche Klasse Fisch es sich bei gegebener Länge handelt. (2 Punkte)

d) Ändert die *Loss Funktion* auf eine beliebige Art ab. Beschreibt (im Idealfall vorher), wie sich das Klassifikationsverhalten ändert. (2 Punkte)