6. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse A
Abgabe: 02.07.2019 23:59

SoSe 2019
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD19]
Do.14–16	CP-03–150	kevin3.schmidt@udo.edu und maximilian.sackel@udo.edu
Fr. 10–12	CP-03–150	tobias.hoinka@udo.edu und noah.biederbeck@udo.edu
Fr. 16–18	CP-03–150	felix.geyer@udo.edu und rune.dominik@udo.edu

Aufgabe 14: *k-NN Klassifikation* **6 P.**

(a) Worauf müssen Sie bei einem k-NN-Algorithmus achten, wenn die Attribute sich stark in ihren Größenordnungen unterscheiden?

(b) Warum bezeichnet man den k-NN als sogenannten „lazy learner“? Wie sind die Laufzeiten für Lern- und Anwendungs-Phase? Wie sind sie im Vergleich zu anderen Algorithmen wie bspw. einer SVM?

(c) Implementieren Sie einen k-NN Algorithmus zur Klassifikation von Ereignissen. Halten sie sich hierbei an die in der beiliegenden Datei `class_structure.py` vorgegebenen Klassenstruktur. Die Methode `predict` soll dabei ein numpy array ausgeben, dass das vorhergesagte Label für jedes Sample enthält. **Vorgehen:** Für jedes zu klassifizierende Ereignis:

1) Berechnung der Abstände zu allen Punkten des Trainingssamples.
2) Bestimmung der k Trainingsevents mit dem kleinsten Abstand (Hinweis: Ermitteln Sie nur die Indizes der Ereignisse, statt das Array an sich zu sortieren).

   Tipp: Die Python-Funktion `numpy.argsort()` ist hilfreich.
3) Bestimmung des Labels, das in diesen Ereignissen am häufigsten vorkommt.

(d) Wenden Sie ihren Algorithmus auf das Neutrino Monte-Carlo von Blatt 5 an. Benutzen Sie die im Moodle zur Verfügung gestellte Datei `NeutrinoMC.hdf5`.

- Nutzen Sie die Attribute `AnzahlHits`, `x` und `y`.
- Setzen Sie $k = 10$.
- Nutzen Sie je 5000 Ereignisse als Trainingsset.
- Das Testset soll aus 20 000 Untergrund- und 10 000 Signalevents bestehen.

Bestimmen Sie Reinheit, Effizienz und Signifikanz.

(e) Was ändert sich, wenn Sie `log10(AnzahlHits)` statt `AnzahlHits` nutzen?

(f) Was ändert sich, wenn Sie $k = 20$ statt $k = 10$ verwenden?

6. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse A
Abgabe: 02.07.2019 23:59

SoSe 2019
Prof. W. Rhode

Aufgabe 15: *Trennende Geraden* **4 P.**

Gegeben seien die Populationen `P_0_10000` und `P_1` aus der Aufgabe „Zwei Populationen“. Nutzen Sie das dort erstellt HDF-File für diese Aufgabe. (Sie finden die Datei ebenfalls im Moodle.) Außerdem seien die Projektionsgeraden

$$g_1(x) = 0 \tag{1}$$
$$g_2(x) = -\frac{3}{4}x \tag{2}$$
$$g_3(x) = -\frac{5}{4}x \quad . \tag{3}$$

gegeben.

(a) Stellen Sie die beiden Populationen zusammen in einem zweidimensionalen Scatterplot dar und zeichnen Sie die drei Projektionsgeraden ein. Im Folgenden werden diese beiden Populationen mit $P0$ und $P1$ bezeichnet.

(b) Projizieren Sie die Punkte aus Population $P0$ und $P1$ jeweils auf die Geraden g_1, g_2 und g_3. Bestimmen und normieren Sie vorher den Projektionsvektor und wählen sie das Vorzeichen so, dass die projizierte Population $P0$ rechts (zu größeren Werten) von $P1$ liegt. Stellen Sie die projizierten Populationen $P0$ und $P1$ für jede Projektion in einem eigenen, eindimensionalen Histogramm dar.

(c) Betrachten Sie $P0$ als Signal und $P1$ als Untergrund. Berechnen Sie die Effizienz und die Reinheit des Signals als Funktion eines Schnittes λ_{cut} in den projizierten Räumen und stellen Sie die Ergebnisse für jede Projektion in einem eigenen Plot dar.

Aufgabe 16: *Naive Bayes: Fußball* **4 P.**

Der Satz von Bayes lautet:

$$P(F|W) = \frac{P(W|F) \cdot P(F)}{P(W)} \tag{4}$$

(a) Beweisen Sie den Satz von Bayes (4) mit Hilfe der Definition der bedingten Wahrscheinlichkeit.

In dieser Aufgabe beschreibt F, ob Fußball gespielt wird oder nicht. W beschreibt die Wetterbedingung, welche durch vier Attribute beschrieben wird. Ihnen steht der Datensatz in Tabelle 1 zur Verfügung.

Attribut	$F =$ ja	$F =$ nein
Wind	schwach(6), stark(3)	schwach(2), stark(3)
Feuchtigkeit	hoch(3), normal(6)	hoch(4), normal(1)
Temperatur	heiß(0), mild(6), kalt(3)	heiß(1), mild(1), kalt(3)
Ausblick	sonnig(2), bewölkt(4), regnerisch(3)	sonnig(1), bewölkt(1), regnerisch(3)

Tabelle 1: Datensatz. (In den Klammern steht wie oft der entsprechene Wert gemessen wurde.)

b) Die Wetterbedingungen des heutigen Tages finden Sie in Tabelle 2. Wie hoch ist die Wahrscheinlichkeit, dass heute Fußball gespielt wird?

Attribut	Wert
Wind	stark
Feuchtigkeit	hoch
Temperatur	kalt
Ausblick	sonnig

Tabelle 2: Die Wetterbedingungen heute.

Tipps:

1. Mit der naiven Annahme, dass die Attribute x_i unabhängig sind, gilt:

$$P(W|F) = \prod_i P(x_i|F)$$

2. Überlegen Sie sich woraus sich die Normierung $P(W)$ zusammen setzt.

c) Nehmen Sie an, Sie sollten nun berechnen, wie hoch die Wahrscheinlichkeit ist morgen Fußball zu spielen (Wetterbedingungen s. Tabelle 3). Welches Problem tritt auf und wie kann man es lösen?

Attribut	Wert
Wind	schwach
Feuchtigkeit	hoch
Temperatur	heiß
Ausblick	sonnig

Tabelle 3: Die Wetterbedingungen morgen.

Aufgabe 17: *Binärer Entscheidungsbaum: Die erste Entscheidung* **6 P.**

Sie haben einen Datensatz wie er in Tabelle 4 gegeben ist. Hierbei ist

- Temperatur: Temperatur in Grad Celsius.
- Wettervorhersage: Wetterqualität (0: schlecht , 1: normal, 2: gut).
- Luftfeuchtigkeit: Luftfeuchtigkeit in Prozent.
- Wind: Aussage, ob es gerade windig ist.
- Fußball: Lohnt es sich Fußball spielen zu gehen?

Hierbei ist das Zielattribut, welches man bestimmen will, die Entscheidung, ob es sich lohnt Fußball spielen zu gehen. In dieser Aufgabe sollen Sie zu diesem Zweck den ersten Schnitt eines *binären* Entscheidungsbaumes nachvollziehen.

(a) Berechnen Sie per Hand die Entropie der Wurzel (des Baumes).

(b) Berechnen Sie per Hand den Informationsgewinn, falls ein Schnitt auf dem Attribut `Wind` durchgeführt wird.

(c) Berechnen Sie für die verbleibenden Attribute den Informationsgewinn in Abhängigkeit von verschiedenen Schnitten und plotten Sie den Informationsgewinn in Abhängigkeit der jeweiligen Schnitte.

(d) Welches Attribut eignet sich am besten zum Trennen der Daten?

Tabelle 4: Datensatz: „Soll ich Fußballspielen gehen?“

Temperatur / °C	Wettervorhersage	Luftfeuchtigkeit / %	Wind	Fußball
29,4	2	85	False	False
26,7	2	90	True	False
28,3	1	78	False	True
21,1	0	96	False	True
20	0	80	False	True
18,3	0	70	True	False
17,8	1	65	True	True
22,2	2	95	False	False
20,6	2	70	False	True
23,9	0	80	False	True
23,9	2	70	True	True
22,2	1	90	True	True
27,2	1	75	False	True
21,7	0	80	True	False