10. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 19.01.2015, 18:00 Uhr

WS 2014/2015
Prof. W. Rhode

Auf diesem Zettel lassen sich die Hälfte der Punkte ohne Programmieren erreichen

Aufgabe 29: *k-NN Klassifikation* — 10 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“?

c) Was ist der Hauptgrund für die vergleichsweise (z.B. gegenüber eines Random Forests) lange Laufzeit eines k-NN Algorithmus?

d) Implementieren Sie einen k-NN Algorithmus zur Klassifikation von Ereignissen. Die Funktion soll hierbei das Trainingssample, die Label des Trainingssamples, die zu klassifizierenden Daten, sowie das k übergeben bekommen. Die Rückgabe sollen die ermittelten Label für die Datenereignisse sein.

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).

   Im EWS finden Sie zwei C++ Programme, `argsort98.cpp` und `argsort11.cpp`, welche die Sortierung der Indizes für C++98 bzw. C++11 exemplarisch darstellt. Für eine Python-Lösung ist die Funktion `numpy.argsort()` hilfreich.
3) Bestimmung des Labels, das in diesen Ereignissen am häufigsten vorkommt.

e) Wenden Sie ihren Algorithmus auf das Neutrino Monte-Carlo von Blatt 7 an. Benutzen Sie die im EWS zur Verfügung gestellte Datei `Blatt7.root`.

- 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.

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

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

Aufgabe 30: *Binärer Entscheidungsbaum: Die erste Entscheidung* **10 P.**

Sie haben einen Datensatz wie er in Tabelle 1 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.

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 1: 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