6. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 05.12.2017, 23:59 Uhr

WS 2017/2018
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD1718]
Di. 10-12	CP-03-150	philipp2.hoffmann@udo.edu und jan.soedingrekso@udo.edu
Di. 16-18	P1-02-110	felix.neubuerger@udo.edu und tobias.hoinka@udo.edu
Di. 16-18	CP-03-150	simone.mender@udo.edu und maximilian.meier@udo.edu

Aufgabe 19: *k-NN Klassifikation* **7 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. einem Random Forest?

c) 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).

   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 3 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?

Aufgabe 20: *Multivariate Regression* **7 P.**

In dieser Aufgabe sollen Sie eine 2 dimensionale multivariate Regression mit *sklearn* durchführen und die Ergebnisse anschaulich darstellen.

a) Erstellen sie ein Dataframe mit 10^5 uniform zwischen 0 und 1 verteilten Zufallszahlen x_1, x_2.

b) Berechnen sie aus diesen Attributen ein drittes Attribut x_3 mit der Funktionsvorschrift:

$$x_3 = 15\sin(4\pi x_1) + 60(x_2 - 0.5)^2$$

Addieren Sie auf diese Zahl eine standardnormalverteilte Zufallszahl, um Rauschen zu simulieren. Das x_3 Attribut ist von nun an Ihr Zielattribut.

c) Teilen Sie das Dataframe in einen Trainings- und Test-Datensatz auf.

d) Wählen Sie einen Random-Forest-Regressor mit 200 Bäumen und trainieren Sie diesen auf dem Trainingsdatensatz um x_3 zu schätzen.

e) Stellen Sie die erstellten Daten und die Vorhersagen des Regressors in einem dreidimensionalen Plot und mehreren 2 dimensionalen Projektionen dar um die Vorhersage mit der Wahrheit zu vergleichen. Geben sie außerdem den *mean-squared-error* der Vorhersage zu den wahren Werten an.

f) Erstellen Sie einen weiteren Datensatz, bei dem x_1 und x_2 uniform verteilte Zufallszahlen zwischen 1 und 2 sind. Was für Probleme können hier auftreten und was ist die Vorhersage des Regressors? Stellen Sie das Ergebnis wie in Aufgabe **e)** dar.

Aufgabe 21: *Binärer Entscheidungsbaum: Die erste Entscheidung* **6 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 (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 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