Prof. Dr. Stephan Trahasch
Betriebliche Kommmunikationssysteme
und IT-Sicherheit

Hochschule Offenburg
offenburg.university

Übungsblatt 2

k-NN und naïve Bayes, 25 Punkte

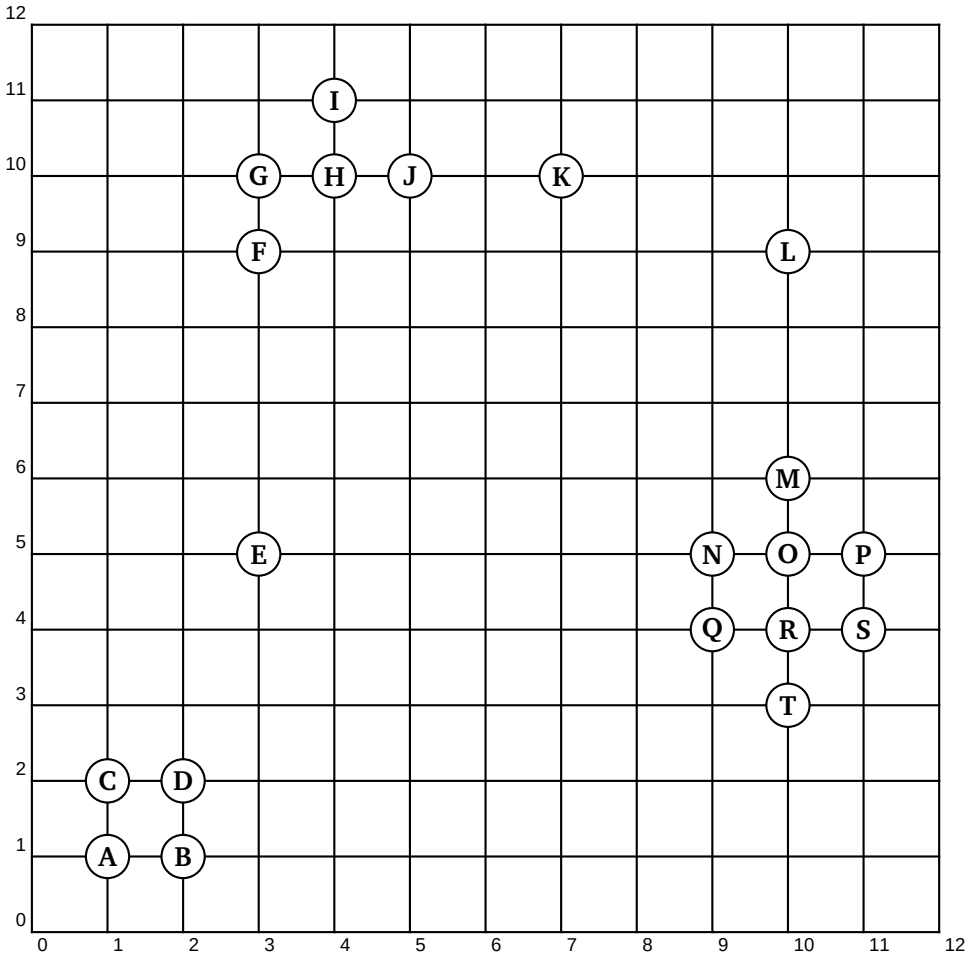
Data Mining

Wintersemester 2016/17

Abgabe: 18.11.2016 9:45 Uhr

Aufgabe 1 Outlier Scores mit kNN 5 Punkte

Gegeben seien folgende 2-dimensionale Daten:



Verwenden Sie als Distanzfunktion zwischen den Punkten die Manhattan-Distanz (L_1-Norm). Berechnen Sie (unter Ausschluss des Anfragepunktes bei der Berechnung der kNN):

- Die kNN-Distanz für $k = 2$ für alle Punkte.
- Die aggregierte kNN-Distanz für $k = 2$ für alle Punkte.

Diskutieren Sie die Wahl von k für diesen Datensatz.

Aufgabe 2 k-NN in R 8 Punkte

Sie untersuchen in dieser Aufgabe den Datensatz vehicle.dat. Informationen dazu finden Sie unter http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29 Dieser umfasst 846 Beobachtungen und 19 Variablen. Ziel ist es, eine Silhouette dem richtigen von 4 Autotypen zuzuordnen. Die Silhouette ist dabei gegeben durch 18 verschiedene Merkmale wie Kompaktheit, Längenverhältnisse, usw. Die Zielvariable heisst Class und umfasst die Typen bus, opel, saab und van. Den Datensatz vehicle.csv finden Sie im Ordner data

Die Datei 02_knn_Student.R enthält das Grundgerüst für die Aufgabe. Integrieren Sie Ihre Lösungen in diese Datei.

Bestimmen Sie ein optimales k und trainieren Sie k-NN-Modelle in folgenden Varianten

a) ohne Normalisierung der Daten

b) mit Normalisierung der Daten

c) mit einer z-Transformation der Daten

Vergleichen Sie die Ergebnisse miteinander und diskutieren Sie das Ergebnis.

Vergleichen Sie das beste k-NN Modell mit dem besten Random Forest aus der vorhergenden Übung.

Aufgabe 3 Satz von Bayes 2 Punkte

In einem Online-Shop wird ein neuer Algorithmus zur Betrugserkennung (Fraud-Detection) in Betrieb genommen, der die Echtheit von Kreditkartenzahlungen prüft. Aus Erfahrung weiß man, dass 15 von 10.000 Kreditkartenzahlungen mit gefälschten Kreditkarten durchgeführt werden. Wenn die Kreditkarte gefälscht ist, gibt dieser Algorithmus mit einer Wahrscheinlichkeit von 0,95 eine Warnmeldung aus, und wenn die Kreditkarte korrekt ist, mit einer Wahrscheinlichkeit von von 0,1.

Wie sicher kann man davon ausgehen, dass die Kreditkarte tatsächlich falsch ist, wenn der Algorithmus einen Warnhinweis ausgibt?

Aufgabe 4 Naive Bayes 5 Punkte

Berechnen Sie mit Hilfe der Naive Bayes Klassifikation für den Neukunden mit den Attributen jung, mittel, ja, schlecht die Klassifikation, ob dieser einen neuen PC kauft oder nicht. Basis sind folgende Daten:

Alter	Einkommen	Student	Kreditwürdigkeit	Kauft PC?
Jung	Hoch	Nein	Schlecht	Nein
Jung	Hoch	Nein	Gut	Nein
Mittelalt	Hoch	Nein	Schlecht	Ja
Senior	Mittel	Nein	Schlecht	Ja
Senior	Niedrig	Ja	Schlecht	Ja
Senior	Niedrig	Ja	Gut	Nein
Mittelalt	Niedrig	Ja	Gut	Ja
Jung	Mittel	Nein	Schlecht	Nein
Jung	Niedrig	Ja	Schlecht	Ja
Senior	Mittel	Ja	Schlecht	Ja
Jung	Mittel	Ja	Gut	Ja
Mittelalt	Mittel	Nein	Gut	Ja
Mittelalt	Hoch	Ja	Schlecht	Ja
Senior	Mittel	Nein	Gut	Nein

Tabelle 1: Kunden

Aufgabe 5 SMS Spam Detection with Naive Bayer 5 Punkte

Die SMS Spam Collection Data Set enthält Daten echter Spam-SMS und normalen Nachrichten, sogenannter Ham. Die Daten finden Sie im Repository im Verzeichnis data.

Lesen Sie die Daten ein und erstellen Sie ein Naive-Bayes-Klassifikator. Überlegen Sie sich, welche Schritte Sie zum Lernen des Naive Bayes durchführen müssen und welche Informationen Sie benötigen. Die Datei 02_Naive_Bayes_Spam_Student.R enthält das Grundgerüst für diese Aufgabe.