Information Mining - Wintersemester 2014
Dipl.-Inform. Vu Tran, LF 139
vtran@is.inf.uni-due.de

Übungsblatt 10

Aufgabe 22: Instanzbasiertes Lernen: k-Nächste-Nachbarn

(a) Erläutere und skizziere, wie mit Hilfe von *k-Nächste-Nachbarn* (k-NN) Instanzen klassifiziert werden können.

(b) k-Nächste-Nachbarn gehört zu den sogenannten „faulen“ Lernverfahren. Was bedeutet das?

(c) Für k-NN wird ein Distanzmaß benötigt, welches den Abstand zwischen zwei Instanzen berechnet. Häufig wird das sogenannte Kosinusmaß eingesetzt. Das Kosinusmaß ist definiert als

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|},$$

wobei $\vec{a}$ und $\vec{b}$ zwei Instanzen in einem Vektorraum sind.

Gegeben seien vier Trainingsinstanzen:

$$\vec{a} = \begin{pmatrix} 0 \\ 3 \\ 5 \end{pmatrix}, \vec{b} = \begin{pmatrix} 3 \\ 3 \\ 8 \end{pmatrix}, \vec{c} = \begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix}, \vec{d} = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}$$

Zu welcher Klasse würde man die Instanz $\vec{x} = (5, 8, 0)^T$ zuordnen, wenn $\vec{a}$ und $\vec{b}$ zur Klasse X und $\vec{c}$ und $\vec{d}$ zur Klasse Y gehören (2-NN-Klassifikation mit Kosinusmaß zur Distanzberechnung)?

Aufgabe 23: Kombination mehrerer Modelle

(a) In der Vorlesung wurden Verfahren zur Kombination mehrerer Modelle vorgestellt. Beschreibe kurz in eigenen Worten die zugrundeliegenden Ideen der nachfolgenden Ansätze: (i) Bagging, (ii) Boosting und (iii) Stacking.

(b) Führe in *RapidMiner* eine Klassifikation der Beispieldaten[1] mit dem Stacking-Verfahren durch und evaluiere das Ergebnis mithilfe einer zehnfachen Kreuzvalidierung.

Verwende `Naive Bayes` und `k-NN` (mit $k = 5$ und Kosinusmaß) als Lernverfahren zum Lernen der Modelle. Das Lernverfahren zum Stacking soll `Decision Tree` sein.

[1] `http://www.is.inf.uni-due.de/courses/im_ws14/uebung/data_a23.csv`