Prof. Dr. Morik
Christian Bockermann

Dortmund, 13. Januar 2009

Übung zur Vorlesung

Maschinelles Lernen

Wintersemester 2008/2009
Blatt 11

Wiederholung

- Welche Eigenschaften erfüllt ein Distanzmaß?
- Was ist ein Ähnlichkeitsmaß? Wie ist der Zusammenhang zum Distanzmaß?

In der Vorlesung wurde das Clustering als Lernaufgabe behandelt und einige Clustering-Verfahren vorgestellt. Dieses Übungsblatt widmet sich in gewisser Weise einer gegenteiligen Aufgabe: der Ausreißer-Analyse (*Outlier-Detection*).

Aufgabe 1

5 Punkte

Es existieren in der Literatur unterschiedliche Definitionen zum Thema *Outlier*. Einen intuitiven Ansatz liefert die folgende Definition:

Definition (ε-Ausreißer): *Sei ein festes $\varepsilon > 0$ und ein $k \in \mathbb{N}$ gegeben. Ein Example $\vec{x}$ ist ein Ausreißer, wenn nicht mehr als k Punkte in der ε-Umgebung von $\vec{x}$ liegen.*

1. Implementieren Sie einen Operator, der für gegebenes ε und gegebenes k die ε-Ausreißer in einer Menge von Beispielen bestimmt. Implementieren Sie Ihren Operator wieder als einfaches Prediction-Modell, um auf diese Weise die Visualisierung in RapidMiner nutzen zu können (binäre Klassifikation).
2. Testen Sie Ihren Algorithmus auf dem Datensatz `clusterData1.csv` aus dem Verzeichnis `mlv-uebung/samples`.
3. Untersuchen Sie dabei unterschiendliche Werte für ε, k. Welche Werte identifizieren Ausreißer in dem Datensatz Ihrer Meinung nach am besten?

Aufgabe 2

5 Punkte

Eine etwas andere Erfassung von Ausreißern liefert die folgende Definition:

Definition (D_n^k-Ausreißer): *Sei $\vec{x}$ ein Beispiel und $D^k(\vec{x})$ der Abstand von $\vec{x}$ zu seinem k-ten Nachbarn (nach Abstand sortiert). Für $k, n \in \mathbb{N}$ ist ein Beispiel $\vec{x}$ ein Ausreißer, wenn nicht mehr als $n-1$ Punkte $\vec{q}_1, \ldots, \vec{q}_{n-1}$ einen größeren D^k-Wert haben ($q_i \neq \vec{x}$).*

1. Implementieren Sie einen Operator, der für gegebene $k, n \in \mathbb{N}$ die D_n^k-Ausreißer in einer Menge von Beispielen findet.
2. Testen Sie Ihren Operator wieder auf dem Datensatz `clusterData1.csv` und vergleichen Sie die Ergebnisse mit Ihren Ergebnissen aus Aufgabe 1. Welche Werte für k, n lieferten gute Resultate?