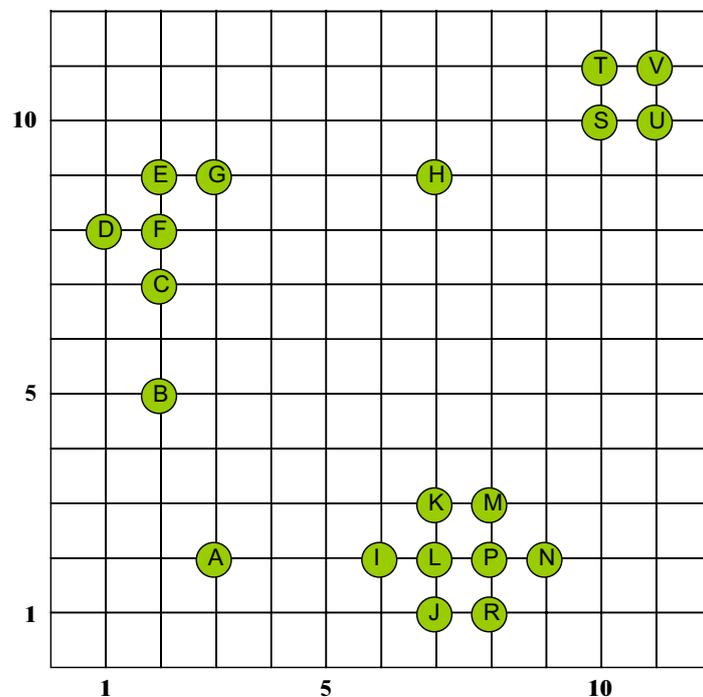


Knowledge Discovery in Databases
 WS 2009/10

Übungsblatt 10: Outlier Detection
 Besprechung am 15.1.2010

Aufgabe 10-1 *Outlier Scores*

Gegeben sei der folgende 2-dimensionale Datensatz:



Verwenden Sie als Distanzfunktion zwischen den Punkten wieder die Manhattan-Distanz (L_1 -Norm)

Berechnen Sie (unter Ausschluss des Anfragepunktes bei der Berechnung der k NN):

- Den LOF-Wert für $k = 2$ für die Punkte H , L und B .
- Den LOF-Wert für $k = 4$ für die Punkte H , L und B .
- Die k NN-Distanz für $k = 2$ für alle Punkte.
- Die k NN-Distanz für $k = 4$ für alle Punkte.

Diskutieren Sie die Wahl von k für diesen Datensatz.

Aufgabe 10-2 *Drei-Sigma-Regel*

In der Statistik findet man oft die sogenannte “Drei-Sigma-Regel”, auch als “68-95-99.7-Regel” bekannt. So spricht man ab einem Wert von mehr als 95% von “schwach signifikant (*)”, ab 99% von “stark signifikant (**)” und ab 99.9% von “sehr stark signifikant (***)”. Bezieht man diese Werte auf eine Normalverteilung, so entspricht dies etwas einer Abweichung von $\pm 2\sigma$ ($\approx 95\%$) bzw. $\pm 3\sigma$ ($\approx 99.7\%$). Eine Abweichung von mehr als 2σ ist also nach dieser Sprechweise “schwach signifikant” und ab 3σ “stark signifikant”.

Jedoch finden diese Regeln in der Statistik normalerweise dann Anwendung, wenn es sich nur ein paar hundert Datensätze handelt. Die Anwendung einer solchen Regel in der automatischen Datenanalyse führt aber zu Problemen:

Berechnen Sie dazu einen einfachen Erwartungswert, wie viele Elemente auf einem Standardnormalverteilten Datensatz von einer Million Werte um mehr als 3σ vom Mittelwert abweicht.

Braucht ein statistisches Verfahren (z.B. EM-Outlier Detection), dass Ausreißer nach einer solchen Verteilung bewertet, daher vielleicht eine Korrektur? Wie kann man dies korrigieren?