Alexander Hinneburg,
hinneburg@informatik.uni-halle.de

19.10.2005

Übung zur Vorlesung „Data Mining in Datenbanken“

Übungsblatt 2

Abgabe am 27.10.2005

Installieren Sie sich den InstantClient von Oracle [1] mit den Paketen *Instant Client Package - Basic* und *Instant Client Package - SQL*Plus* in Ihrem home-Verzeichnis in der Uni[2]. Wählen Sie die richtige Plattform, z.B. für turing.informatik.halle.de brauchen Sie Solaris64. Die Installation besteht im Download einer Zip-Datei für jedes Paket, auspacken in das gleiche Verzeichnis, fertig. Zum Entpacken und starten führen Sie folgende Kommandos aus:

```
unzip instantclient-basic-solaris64-10.2.0.1-20050828.zip
unzip instantclient-sqlplus-solaris64-10.2.0.1-20050828.zip
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:./
cd instantclient_10_2
./sqlplus username/password@mozart.informatik.uni-halle.de/lxdb3
```

Unter Windows wird SQLplus auch von der Kommandozeile aus gestartet:

```
C:\OracleInstantClient>sqlplus username/password@mozart.informatik.uni-halle.de/lxdb3
```

Für *nutzername* und *password* müssen Sie Ihre Nutzerdaten eintragen. Sie können dann mit SQL auf die Tabelle IRIS mit den Irisdaten zugreifen.

Aufgabe 1 Entwickeln Sie ein SQL Statement zur Berechnung eines ein-dimensionalen Histogramms mit gleichgroßen Intervallen zwischen Min/Max der entsprechenden Dimension. Wählen Sie die Anzahl der Intervalle nach der Regel von Sturge $k = 1 + \log_2 n$, wobei n die Anzahl der Datenwerte ist.

Berechnen Sie mit Hilfe dieser Anfragen die Histogramme für petalwidth, petallength, sepalwidth, sepallength. Zeichnen Sie die Histogramme und beschriften Sie die Achsen, z.B. mit Gnuplot `http://www.gnuplot.info/` und geben Sie die vier SQL-Statements an. Sie bekommen Zusatzpunkte, wenn Sie die Werte für Min/Max und k nicht hart in die Statements hinein kodieren.

Um das Ergebnis der SQL Anfragen in eine Datei zu exportieren, verwenden Sie das SQLplus Kommando spool. Die Beschreibungen zu Oracle finden Sie z.B. unter `http://mozart.informatik.uni-halle.de/oracle/DB10g/index.htm` in der Büchern (Reiter Books) *SQL Reference* und *SQL*Plus User's Guide and Reference*. Diese Seiten sind nur aus dem Uni-Netz verfügbar.

Aufgabe 2 Was ist neben der Datenverteilung für die Form eines Histogramm ausschlaggebend? D.h. stellen Sie sich vor, dass die Datenwerte fest sind, wie können Sie trotzdem das Histogramm ändern. Nennen Sie zwei wesentliche Einflußfaktoren.

---

[1] `http://www.oracle.com/technology/tech/oci/instantclient/instantclient.html`

[2] Da das Rechenzentrum die notwendigen Ports nicht freigegben hat, ist eine Direktverbindung von außerhalb der Uni zum DB-Server mozart.informatik.uni-halle.de nicht möglich.

Aufgabe 3 Manchmal sind kumulative Histogramme anstelle von normalen Histogrammen gefragt. Machen Sie einen Vorschlag, um ein kumulatives Histogramm mit einem SQL Statement zu berechnen. Was sind die Schwierigkeiten? Studieren Sie dazu das analytische Beispiel unter `http://mozart.informatik.uni-halle.de/oracle/DB10g/server.101/b10759/functions148.htm#sthref1800` unten. Sie bekommen Zusatzpunkte, wenn Sie ein funktionierendes SQL Statement zur Berechnung eines kumulativen Histogramms von petalwidth angeben können.