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

17.11.2005

Übung zur Vorlesung „Data Mining in Datenbanken“

Übungsblatt 5[1]

Auswahl der Attribute
Abgabe am 24.11.2005

In der letzten Übung haben Sie eine Datenmenge erzeugt, die wenig Instanzen und sehr viele Attribute hat. Für den Naive Bayes Klassifikator stellt eine sehr große Attributmenge ein Problem dar, weil die Annahme gemacht wird, daß alle Attribute von einander unabhängig sind.
Nutzen Sie die in der letzten Übung erstellten normalisierten Daten. Einige Schritte werden einfacher, wenn die Gene die Zeilen bilden und die Patienten die Spalten sind.

Aufgabe 1 Nutzen Sie für die folgenden Schritte nur die Trainingsdaten. In den Daten gibt es zwei Klassen `ALL` und `AML`. Für ein Gen x sei $\bar{x}_1$ der durchschnittliche Expressionswert, σ_1^2 die Varianz und n_1 die Anzahl der Instanzen der Klasse `ALL` und analog $\bar{x}_2$, σ_2^2, n_2 die entsprechende Werte für `AML`. Die Varianz ist wie folgt definiert $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ läßt sich aber praktischer so berechnen

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Der Signal to Noise Ratio für `ALL` ist definiert als

$$S2N(x) = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_1 + \sigma_2}$$

(beachten Sie das fehlende Quadrat bei σ) und der T-Wert für `ALL` ist definiert als

$$T(x) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Die Werte für `AML` sind analog nur die Indizes sind vertauscht.

1. Schreiben Sie ein Programm oder Skript, um für jede Klasse die Gene mit den 50 höchsten $S2N$ bzw. T-Werten zu finden. Geben Sie das Skript an und die vier Genlisten.
2. Was ist die Beziehung der $S2N$ bzw. die T-Werte der verschiedenen Klassen `ALL` und `AML`? Gegeben Sie die Gleichungen oder Ungleichungen an. Gelten die Beziehungen noch, wenn es mehr als zwei Klassen gibt?
3. (a) Geben Sie die Gene an, die in den Top-50 Listen für $S2N$ in beiden Klassen gemeinsam sind. (b) Wiederholen Sie (a) mit Top-3 Listen. (c) Wiederholen Sie (a) und (b) mit T-Werten statt $S2N$. Geben Sie die Kommandos/Skripte/Programme und die Listen an.

[1] Achten Sie auch auf die Form Ihrer Lösungen, z.B. daß jedes Bild eine Bildunterschrift hat. Bitte geben Sie keine losen Blätter ab. Um die Übungen in angemessener Zeit zu diskutieren, bereiten Sie neben Ihrer Lösung ein kleine Präsentation vor, mit der Sie dann Ihre Ergebnisse zeigen können.

4. Trainieren Sie den NaivenBayesSimple Klassifikator für die Datenmenge (a) bestehend den Attributen aus dem Schnitt der Top-50 Listen für $S2N$ und (b) wie (a) aber hier die Top-50 Listen der T-Werte verwenden. Nutzen Sie die entsprechenden Testmengen um die Ergebnisse zu validieren. Beschreiben und diskutieren Sie Ihre Ergebnisse.

Aufgabe 2 Zeigen Sie formal die Mehrschritteigenschaft der Entropie

$$entropie(p, q, r) = entropie(p, q + r) + (q + r) \cdot entropie(q/(q + r), r/(q + r))$$

. Siehe Folien 128, 134 und 135.

Aufgabe 3 Erstellen Sie ein Konzept um den A-priory Algorithmus zur Berechnung der häufigen Item-Mengen zu implementieren. Entwickeln Sie Pseudo-Kode unter der Annahme, daß Sie einen Hash (oder ähnliche Datenstruktur zur effizienten Verwaltung von Schlüssel-Wert Paaren) und Sortieren nutzen können. Die Eingabe soll als Transaktionsliste erfolgen; in einer Zeile stehen nur die Items, die in der Transaktion enthalten sind (Einsen). Der Wert der minimalen Überdeckung ist ein Parameter des Algorithmus. Die häufigen Item-Mengen sollen in keiner speziellen Reihenfolge ausgegeben werden.