Übungen Expressionsdatenanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 2

Aufgabe 2.1 (2 Punkte)

Beweisen Sie, dass jenes μ, welches die Summe $\sum_{i=1}^{N}(x_i - \mu)^2$ minimiert, gleich dem arithmetischen Mittel der x_i ist.

Aufgabe 2.2 (18 Punkte)

Der Datensatz `arab` enthält normierte Expressionsdaten von 1.518 Genen und 72 Experimenten. Um die Resistenz von *Arabidopsis thaliana* gegen das Pathogen *Phytophtora infestans* zu erforschen, wurden Expressionsdaten von *Arabidopsis thaliana* 6, 12 und 24 Stunden nach Befall mit *Phytophtora infestans* erhoben. Zu jedem Zeitpunkt wurden je 3 Hybridisierungen mit befallenen und nicht befallenen Proben durchgeführt. Damit lagen 18 Messwerte pro Gen für insgesamt 22.810 Gene vor. Diese Rohdaten wurden mit 8 verschiedenen Algorithmen normiert, und die resultierenden 144 normierten Expressionswerte pro Gen wurden genutzt, um 1.518 differentiell exprimierte Gene zu identifizieren. Der Datensatz `arab` enthält die zu diesen 1.518 Genen gehörigen 72 normierten Expressionswerte der befallenen Proben.

(a) Nutzen Sie die fünf in der Vorlesung behandelten hierarchischen Clusteralgorithmen zur Clusterung der 1.518 Gene unter Nutzung des Euklidischen Abstandes. D. h. berechnen Sie mit jedem der fünf Clusteralgorithmen genau $K = 2, \ldots, 5$ Cluster, indem Sie die drei agglomerativen Clusteralgorithmen $K - 1$ Schritte vor dem Ende und die beiden divisiven Clusteralgorithmen nach $K - 1$ Schritten stoppen. Berechnen Sie in jedem der fünf Fälle (und für jeden der vier Werte von K) die *Intra-Cluster-Varianz* $W(\underline{k})$. Welchen der fünf Algorithmen würden Sie bevorzugen?

(b) Wenden Sie die vier in der Vorlesung eingeführten Varianten (1a, 1b, 2a, 2b) des k-means Algorithmus auf diesen Datensatz an, und führen Sie für $K = 2, \ldots, 5$ die folgenden beiden Studien durch.

(i) Wählen Sie in Varianten 1a uns 2a jeweils den k-ten Datenpunkt als initiales Clusterzentrum von Cluster $k = 1, \ldots, K$, und ordnen Sie in Varianten 1b und 2b jeden Datenpunkt $i = 1, \ldots, 1518$ dem Cluster $k = i \bmod K$ zu. Berechnen Sie für jede der vier Varianten in jedem Iterationsschritt $W(\underline{k})$, und stellen Sie die Funktionen $W(\underline{k})$ grafisch dar. Fällt $W(\underline{k})$ monoton? In wie vielen Iterationsschritten M wächst $W(\underline{k})$? Fassen Sie M, die erreichte

Intra-Cluster-Varianz $W(\underline{k})$ und die dafür benötigte Anzahl der Iterationen tabellarisch zusammen. Welche der vier Varianten ist für diesen Datensatz und die hier genutzte Initialisierung am besten geeignet?

(ii) Wählen Sie in Varianten 1a uns 2a zufällig (von einer Gleichverteilung gezogen) je einen der 1.518 Datenpunkte als initiales Clusterzentrum von Cluster $k = 1, \ldots, K$, und ordnen Sie in Varianten 1b und 2b jeden Datenpunkt $i = 1, \ldots, 1518$ zufällig (von einer Gleichverteilung gezogen) genau einem der Cluster $k = 1, \ldots, K$ zu. Initialisieren Sie jede der vier Varianten des k-means Algorithmus je 10^3-mal, und berechnen Sie jeweils die erreichte *Intra-Cluster-Varianz* $W(\underline{k})$. Stellen Sie die sechs *scatter plots* der erreichten $W(\underline{k})$ (1a vs. 1b, 1a vs. 2a, 1a vs. 2b, 1b vs. 2a, 1b vs. 2b, 2a vs. 2b), die vier Histogramme der erreichten $W(\underline{k})$ und die vier Histogramme der benötigten Anzahl der Iterationen grafisch dar. Welche der vier Varianten ist für diesen Datensatz am besten geeignet? Zählen Sie außerdem, in wie vielen Iterationsschritten M die *Intra-Cluster-Varianz* $W(\underline{k})$ wächst, und stellen Sie die vier Histogramme der jeweils 10^3 Werte von M grafisch dar.

Transponieren Sie nun die Datenmatrix und wiederholen Sie gesamte Aufgabe, d. h. clustern Sie nun die Experimente und nicht die Gene. Stellen Sie in diesem Fall die durch die fünf hierarchischen Clusteralgorithmen berechneten Dendrogramme grafisch dar.

Abgabetermin: 15. Mai