Einführung in die Bioinformatik

Prof. Dr. Enno Ohlebusch, Prof. Dr. Hans Kestler, Dr. Karlheinz Holzmann,
Tobias Badura

WS 16/17

Übungsblatt 5

Abgabe bis 08.02.2017, 12:00. Lösungen bitte elektronisch (an tobias.badura@uni-ulm.de) mit Namen in Text und Quelltextdateien abgeben.

1. Aufgabe (6): Preprocessing

Für diese Aufgaben wird der `Golub`-Leukämie-Datensatz (ALL versus AML) von der Vorlesungswebseite benötigt. Lesen Sie diesen mit `read.table("rgolub.train",check.names=FALSE)` in die Variable X ein. Die Zeilen enthalten Gene und die Spalten die verschiedenen Fälle (ALL, AML). Konvertieren Sie den eingelesenen data.frame in eine matrix mit `as.matrix()`.

(a) Begrenzen Sie die Expressionswerte elementweise auf das Intervall $[100, 16000]$ d.h. Werte < 100 werden auf 100 gesetzt und Werte > 16000 werden auf 16000 gesetzt.

(b) Filtern Sie alle Gene heraus, so daß für jedes übriggebliebene Gen gilt $max/min > 5$ und $max - min > 500$ (über alle Samples). Geben Sie die Anzahl der übriggebliebenen Gene aus.
Hinweis:
Schreiben Sie eine Funktion (oder eine Lambda-Funktion), die für eine gegebene Zeile einen booleschen Wert zurückgibt und wenden sie `apply` an.

(c) Normalisieren Sie den Datensatz mit der Lowess Regression.
Hinweis: Verwenden Sie hierzu `lowess` aus dem `stats` Paket.

(d) Erstellen Sie einen MA-Plot für die Ausgangsdaten und die normalisierten Daten.

(e) Logarithmieren Sie alle normierten Expressionswerte (Logarithmus zur Basis 2).

(f) Plotten Sie eine Häufigkeitsverteilung (Histogramm) der Genexpressionswerte aus Aufgabe (e).

2. Aufgabe (2): Multiples Testen

Laden Sie den Genexpressionsdatensatz `multTest.txt` von der Vorlesungswebseite herunter (Zeilen = Konditionen, Spalten = Gene, letzte Spalte = Klassenlabel).

(a) Geben Sie die statistisch signifikanten Gene zwischen kranker und gesunder Gruppe auf einem Signifikanzniveau von 5% an. Begründen Sie die Wahl der Teststatistik.

(b) Korrigieren Sie für multiples Testen. Geben Sie die berechneten Gene unter Verwendung der Bonferroni-Korrektur sowie der Holm-Prozedur an.

3. Aufgabe (6): Partitionierende Clusteranalyse

Implementieren Sie den *k-means* Clusteralgorithmus in R. Verwenden Sie die Euklid'sche Distanz als Abstandsmaß. Als anfängliche Clusterzentren sollen zufällig Punkte aus der gegebene Datenmenge gezogen werden. Testen Sie den Algorithmus mit dem Golub-Datensatz von der Vorlesungswebseite (`golub50.test`). Führen unterschiedliche Initialisierungen der Clusterzentren zu unterschiedlichen Ergebnissen?

4. Aufgabe (5): Lowess

Auf der Vorlesungsseite finden Sie unter Zusatzmaterial das Paper 'Robust Locally Weighted Regression and Smoothing Scatterplots'. Beschreiben Sie kurz (max 1/2 Seite) den dort vorgestellten Algorithmus.

5. Aufgabe (5): Permutations-Test

Führen Sie einen Permutations-Test auf Gleichheit der Mittelwerte in beiden Gruppen für den folgenden Datensatz durch:

$$x = (0.66, 0.51, 1.12, 0.83, 0.91, 0.50)$$
$$y = (0.41, 0.57, -0.17, 0.50, 0.22, 0.71) \quad .$$

Implementieren Sie hierzu ein R-Skript.

Hinweis: Erzeugen Sie mit der `sample()` Funktion eine Permutation der Indizes $1 \ldots 12$ und verwenden Sie die ersten 6 und letzten 6 Positionen für die Berechnung der Test-Statistik (=Abstand der Mittelwerte). Orientieren Sie sich an dem Bootstrapping-Beispiel aus der Vorlesung.

Hinweis:

Es wird noch ein letztes Übungsblatt geben, welches bis zum 15.02.17 bearbeitet werden muss, dieses wird spätestens bis zum 08.02. auf der Vorlesungsseite erhältlich sein.