Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

http://uni-leipzig.de/~zuber/teaching/ws11/r-kurs/

Auch dieses Übungsblatt ist dem Datensatz aus der Veröffentlichung *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression* von Golub et al. gewidmet.

- Genexpressionsdaten mit $n = 38$ Beobachtungen und $p = 3051$ Genen
- Faktorvariable `golub.cl` beschreibt welche Form von Leukämie ($k = 2$) bei der entsprechenden Beobachtungen vorliegt
  - ALL: acute lymphoblastic leukemia ($n_0 = 27$)
  - AML: acute myeloid leukemia ($n_1 = 11$)
- Die Matrix `golub.gnames` enthält Information zu den Bezeichnungen der beobachteten Gene
- Die Daten sind schon praeprozessiert und zu finden als `RData`-File im Netzwerkordner unter L:\R-Kurs oder im `R`-Paket `multtest` unter *Data*

1 Aufgabe: Klassifikation

(a) Teilen Sie die Golubdaten zufällig einen Trainings- und einen Testdatensatz. Der Trainingsdatensatz soll dabei 22 Stichproben enthalten, der Testdatensatz die restlichen 16. (HINWEIS: verwenden Sie dazu die `R` Funktionen `runif`, `round` und `unique`)

(b) Berechnen Sie den Fold-Change auf dem Trainingsdatensatz.

(c) Berechnen Sie den t-score auf dem Trainingsdatensatz als Fold-Change / Standardabweichung.

(d) Erstellen Sie eine Rangliste in der die Gene nach der absoluten Größe des t-scores geordnet sind.

(e) Nutzen Sie die top 20 Gene dieser Rangliste, um mittels der Funktion `lda` des Pakets `MASS` eine linearen Klassifikator zu konstruieren und wenden Sie ihn auf den Testdatensatz an.

(f) Berechnen Sie den Vorhersagefehler. Bestimmen Sie außerdem wieviel Prozent der ALL bzw. AML Fälle falsch vorhergesagt werden.

Bestimmen Sie wie in Teilaufgabe (a) eine neue Aufteilung von Trainungs- und Testdatensatz und bestimmen Sie erneut die Fehlerraten. Was fällt Ihnen auf?