Wintersemester 2006/2007 Abgabe am 1. Februar 2007
Prof. Dr. Stefan Posch, Dipl.-Bioinf. Jan Grau

14. Übung „Angewandte Bioinformatik mit Perl und R“

1. Laden Sie wieder die `golub`-Daten sowie das Paket `class`. In diesem Paket gibt es eine Implementierung des kNN-Klassifikators (namens `knn`).

   Trainieren Sie diesen Klassifikator jeweils mit dem Trainingsdatensatz von Golub und testen Sie ihn am Testdatensatz.

   Testen Sie unterschiedliche Untermengen von Genen (s. letze Übungen) und variieren Sie auch das k des kNN-Klassifikators

   Hinweis: `knn` erwartet wie die meisten Klassifikarren in R je Zeile einer Datenmatrix die Proben/samples. Sie können z.B. mit `t(exprs(golubTrain))` die Expressionswerte entsprechend transponieren.

2. Laden Sie nun zusätzlich das Paket `e1071`. Dieses enthält eine support vector machine, die mit `svm` trainiert werden kann. Die resultierende SVM kann in der Funktion `predict` auf diesselben oder andere Daten angewendet werden. Nutzen Sie den linearen kernel (`kernel=linear"`).

   Verwenden Sie wiederum unterschiedliche Untermengen der Gene, trainieren sie die SVM auf den Trainingsdaten und wenden Sie sie dann sowohl auf Trainings- wie Testdaten an. Vergleichen Sie wiederum die Anzahl der Klassifikationsfehler.

3. Die `svm` verfügt auch über eine “eingebaute” cross validation (Argument `cross`). Die Funktion `summary` angewendet auf das Ergebnis von `svm` gibt den Prozentsatz korrekt klassifizierter Proben je Iteration der cross validation aus. Testen Sie!