Prepare the solution for *one* of the following assignments and mail it to `alexander.ploner@ki.se`. A valid solution includes a) a short report containing the graphs you feel are relevant, together with a short explanation what they show, b) an R script that generates all graphs and figures used in the report, and (possibly) c) any extra data files that you need to create in order to run the R script.
Data files referred to in the assignments can be downloaded from the course web page at `www.meb.ki.se/~aleplo/R2007`.

1. The file `Golub.RData` contains expression and phenotypic data for 72 lymphoma patients suffering from ALL or AML, see PubMed 10521349 for details.

   (a) Preprocess the expression data: values below 100 or above 16,000 to be replaced by the thresholds, $\log 2$ of the thresholded values to be taken.

   (b) Try to group patients based on their expression profiles. Is the grouping you find stable when you vary the clustering procedure and/or the distance measure? Is the grouping related to any phenotypic variable?

   (c) Build a KNN predictor that classifies a patient as being either ALL or AML, based on their gene expression. Choose the predictor that performs best on the training set over different sizes of neighborhoods ($k = 1, 3, 5, 7$) and different number of selected features ($n = 10, 50, 100, 200$), using leave-1-out cross validation. Estimate the generalization error of the best predictor on the test set.

   (d) Use the function `svm` in package `e1071` to build a support vector machine that classifies patients as ALL or AML; proceed as above for the KNN, but note that the SVM with the default kernel (`radial`) can be optimized over two parameters, `gamma` and `cost`.
   Suggestion: function `tune.svm` may be helpful.

2. Install Bioconductor on your machine. Download the original `.cel` files and patient data for the ALL data set used in the lecture, available at `http://bioconductor.org/docs/papers/2003/Chiaretti/Chiaretti`. Re-do the analysis using Bioconductor.

   (a) Read in the data using `read.affybatch` in package `affy`.

   (b) Read in the patient data using `read.table`, and store it in the `exprSet` object created in the previous step.

   (c) Compute GCRMA expression values using `gcrma` in the package of the same name.

   (d) Cluster expression values as described in the lecture; use function `genefilter` from the package of the same name to select features.

   (e) Build a KNN predictor as demonstrated in the lecture, but use the function `knnB` of package `MLInterfaces`.