Assignment 2. Due Feb 1.

Correction -- in red.

(a)

In the above example, the Kendall rank correlation was about 0.38. Describe a simple change in the simulation to increase this correlation. Also to decrease it.

Using the same model, generate a test data set with $n = 10^4$ samples. Compare the confusion matricies for the predictions using the model previously fit to the training data.

(b)

The Scenario 2 simulated data for training is available on our website in the file: Scenario2Train.csv. Using R, plot this data along with the decision boundaries for linear and logistic regression.

(c)

Explain briefly why the Regression error rate of $\eta = 27.3\ \%$ is nearly equal to the error rate of k-NN with $k = 67$ as shown on the plot.

(d)

■ Data

Data source: Nathalie Pochet, Frank De Smet, Johan A.K. Suykens and Bart L.R. De Moor (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics Advance Access published July 1, 2004. http://homes.esat.kuleuven.be/~npochet/Bioinformatics/.

The R workspace **`Singh.Rdata`** is available on our webpage. It contains a list variable **`Singh`** with named elements, **`X, y, Xt, yt`**, which are respectively the training design matrix, training classes, test design matrix and test classes.

■ Description

High-quality expression profiles were successfully derived from 52 prostate tumors and 50 nontumor prostate samples from patients undergoing surgery. Oligonucleotide microarrays containing probes for approximately 12600 genes and ESTs. Since prostate tumors are among the most heterogeneous of cancers, both histologically and clinically, the goal here is to classify tumor and nontumour samples. The training set consists of 102 prostate tissues of which 50 are normal and 52 tumor samples. The test set consists of 34 tissues of which 9 are normal and 25 tumor samples. The number of gene expression levels is 12600.

Prediction Problem

Use regression, logisitic regression and kNN to fit a model to the training data. Find the mis-classification rates and the confusion matrix. Then compare using the fitted model to predict on the test data.

In many applied statistics methods, such as ridge regression we standardize the input variables. But with microarray data row standardization is often done to reduce the variability between subjects. Experiment with row and column standardization.

Since $p > n$, you will need need to do gene selection when using regression. Compute the pairwise absolute value of the t-score for each gene and select the largest M scores, where $M < n$.

Microarrays References

This is optional reading for more background but it is not needed for the assignment.

Speed (2000). Statistical analysis of gene expression microarray data

Florian Hahne (2008). Bioconductor case studies