Statistical Machine Learning for Bioinformatics	**Homework 2**
Module II: Comp/Stat 470 Spring 2009	**9 March 2009**

This homework is due March 17, 2009.

There are two problems in this homework. In the first, you will do a comparison of two classification techniques: support vector machines and k-nearest neighbors on the Broad prostate cancer microarray data set. You will also experiment with several feature selection methods. In the next (optional, extra-credit) problem, you will build a high scoring Bayesian network for T-cell signaling from a large proteomic data set obtained from flow cytometry.

1 Molecular classification of prostate cancer by SVMs and k-NNs (100 points)

First, some questions to review your understanding of support vector machines and k-nearest neighbors.

a. Precisely state the optimization problem that the hard margin SVM solves, given a labeled training set $S = \{x^{(i)}, y^{(i)} \mid i = 1 \ldots m\}$, with $x^{(i)} \in \Re^d$ and $y^{(i)} \in \{-1, 1\}$. Why might this version of SVM be inappropriate for real data?

b. Training an SVM produces a set of weights, one per example. Explain briefly how these weights arise from solving the optimization problem given above, and explain how they determine a linear classifier. How do we interpret a weight that is 0, that is non-zero but small, or that is non-zero and large?

c. What are slack variables, are how are they are used to define the soft-margin version of the SVM optimization problem?

d. Why are feature selection techniques crucial for analysis of microarray data sets?

e. What are the fundamental differences between k-nearest neighbor classifiers and SVMs? Which technique would be expect to perform better on microarray data sets such as the Broad prostate cancer data set? Why?

The Broad data is in `broaddata.mat`, the genes are in `genes.mat`, and the labels are in `labels.mat`. You can load these directly into Matlab with the `load` command. For all the experiments below, you will use 5-fold cross validation for training and testing. The Matlab function `crossvalind` splits the 102 data points in the Broad data set into a train and a test set. The `classperf` function analyzes the accuracy of the classifier on a specified test set. Use the `svmclassify` and `knnclassify` functions to classify using SVMs and k-NNs. Feature selection should be done with the `rankfeatures` function. Choose the default `ttest` option and the `roc` option in the rankfeatures function.

a. Train linear SVM classifiers using `rankfeatures` with two feature selection techniques: ttest and roc. Use the top 5, 10, 50, 100 genes. Perform feature selection first, and classify using a hard-margin linear SVM on the reduced feature space. In a cross-validated setting, make sure you chain the feature selection and classification steps, one for each fold. For each feature

set report the class loss, the specificity and sensitivity as well as area under the ROC curve. Present your results in an easy-to-interpret table (2 feature selection techniques 4 sets of feature sizes = 8 rows; and 4 columns (class-loss, sensitivity, specificity, and ROC)). Print the top 10 genes selected by ttest and roc. Is there any overlap between them? Which feature selection method does better? Why? How important is feature selection for this dataset?

b. Repeat the eight experiments in the first part (2 methods x 4 feature set sizes) above on a normalized version of the Broad data set (use the `AutoScale` feature of `svmtrain`). Does normalization change the results? Why or why not?

c. Now rerun the SVM experiments using a radial basis function kernel. You have to choose a width for the kernel. Again, for each, report the class loss, specificity, sensitivity and area under the ROC curve. Did the use of radial basis kernels with the SVM improve performance or lead to overfitting?

d. Replace the SVM classifier in the previous experiments by a kNN classifier with k = 1, 3,5 (use `knnclassify`). Choose a distance measure for the knn. Present your results for each feature selection method (8 experiments for each choice of k). How does the performance of the SVMs compare with the kNNs? Which technique does better and why?

2 Bayesian networks for T cell signaling (extra credit) (100 points)

Classifiers help us identify the genes that are differentially expressed in disease. We can then find an mRNA signature for the disease which is useful during diagnosis. However, to treat disease we need to understand how these genes disrupt fundamental cellular processes. We therefore need to gather data on proteins expressed in the cell as well as their states (phosphorylated or not), to derive a process level explanation. Bayesian networks are one way of representing protein-protein interactions in a cell. It is possible to learn the structure of such an interaction network from flow cytometry data. In this exercise you will learn a Bayesian network from the data set in ofcs3cd28.xls obtainable from `http://www.sciencemag.org/cgi/content/full/sci;308/5721/523/DC1`. It has data on proteins praf, pmek, plcg, PIP2, PIP3, p44/42, pakts473, PKA, PKC, P38, and pjnk. Protein names that begin3 with lower-case p are phosphorylated proteins. The data is obtained from the Sachs' et. al. 2005 Science paper available from Entrez Pubmed by typing Sachs K at the window, and choosing the first link (to a Science April 2005 article). To learn the Bayesian network, use Kevin Murphy's Bayes Net Toolbox (BNT) for Matlab (downloadable from http://bnt.sourceforge.net/). The tutorial at http://bnt.sourceforge.net/usage.html is the best resource for learning how to use the package. Use the structure learning functions to learn the network on the 11 proteins; MCMC is a good choice of method. You will need to discretize the data. Use the Matlab function prctile to discretize each protein level into three categories of low, normal and high. What is the likelihood score of the best network you are able to learn? How does it compare to the network learned in the Sachs' et. al. paper? If there is a difference between your network and theirs, comment on the reasons for the difference.