

This lab session is designed to get you familiar with the state-of-the-art machine learning technique in classification. You will learn how to apply Support Vector Machines (SVM) for tumor classification based on Microarray gene expression data. The SVM package (written in MATLAB) from LIBSVM will be used. Three data sets, including colon, leukemia, and prostate are used in this lab session.

The whole process includes three steps:

- Partition the whole data set into training and test set with a fixed ratio (2/3 of the data used in the training set and 1/3 for testing). A MATLAB function “splitting_ratio” is provided for the partition.
- Build a SVM model for the training set using a specific value of the regularization parameter C, which controls the tradeoff between the maximization of the margin and the minimization of the number of errors made. A MATLAB function "svmtrain" is provided for building the model.
- Classify all data in the test set based on the model built from the training set and report the classification accuracy. A MATLAB function “svmpredict” is provided for prediction.

Part I

A MATLAB function “main”, which combines all three steps, is provided. The input to “main” is the ID of the data set between 1 and 3, and the output is the classification accuracy.

Run “main” ten times on each of the three data sets (the partition into training and test set will be different for different runs) and keep all accuracies as follows:

dataset = 1; (try the other two data sets using 2 and 3)

for i = 1:10

accuracy(i) = main(dataset);

end;

Report the mean accuracy and the standard deviation for each of the three data sets: mean(accuracy), std(accuracy). Summarize your findings.

Part II

A MATLAB function “main_C” is provided. It outputs a set of classification accuracies using different values of the regularization parameter C. In many cases, the value of C has a significant impact on the performance of SVM.

Run “main_C” on each of the three data sets and keep all accuracies as follows:

dataset = 1; (try the other two data sets using 2 and 3)

accuracy = main_C(dataset); (You may run this many times for each data set)

Report all accuracies for different values of C for each of the three data sets. Summarize your main findings from this experiment.