Materials Informatics – Fall 2018

Computer Project 2

Due on: Oct 30

Assignment 1: In this part you will simulate two QSPRs X_1 and X_2 and a property Y, using a Gaussian model, where the prior probabilities are $P(Y=0) = P(Y=1) = \frac{1}{2}$, the means are at the points $\boldsymbol{\mu}_0 = (0,0)$ and $\boldsymbol{\mu}_1 = (1,1)$, and common covariance matrix

$$\Sigma_0 = \Sigma_1 = \sigma^2 \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

Note that this is a homoskedastic Gaussian model (so that the optimal decision boundary is a hyperplane).

(a) Generate a large number (e.g., $M = 1000$) synthetic training data sets for each sample size $n = 20$ to $n = 100$, in steps of 10, with $\sigma = 1$. For each training set and sample size, generate a corresponding independent test set of size $L = 400$. Plot the **average** classification errors of the LDA, 3NN, and linear SVM classification rules, estimated with the test sets, as a function of n on the same graph. Repeat for $\sigma = 2$. Explain what you see.

(b) Using the same synthetic training data generated in part (a), plot the **average** apparent error, leave-one-out, and 5-fold cross-validation error error estimates for the LDA, 3NN, and linear SVM classification rules as a function of n. Generate two plots, one for each classification rule. In each plot, display the average classification error, computed in part (a), and the average error estimates. Repeat for $\sigma = 2$. Explain what you see in terms of error estimation bias. Which error estimators are optimistic, and which are pessimistic? Which error estimator would you would choose for each classification rule, based on these results?

Assignment 2: Here we are going to use the same SFE material data set used in Project 1, categorizing the data into two classes, low (SFE ≤ 35) and high (SFE ≥ 45). We are going to use fixed training and testing data sets, which should be downloaded from e-campus.

We are going to search for feature sets that best discriminate low stacking fault energy from high, using the provided training data, for a given classification rule and different feature selection methods.

We will consider LDA and 3NN as classification rules, and wrapper feature selection, with the apparent error estimate of the designed classifier as the criterion. Finally, we will employ two simple feature selection methods: ehaustive search (for 1 to 5 variables) and sequential forward search (for 1 to 5 variables). If two candidate feature sets have the same minimum apparent error, pick the one with the smallest indices (in "dictionary" order): compare the smallest index, if tied, compare the second smallest one, etc.

Each person will submit a table with the feature sets found. For each row of the table (variable set found), the corresponding error estimate and the test-set estimate (using the test set provided) should be indicated. Intepret the results. In particular, here are examples of questions you should address:

- How do you compare the results against each other and against the results obtained with the simple filter feature selection used in Project 1?
- How do you compare the error estimators and feature selection methods used based on the variable sets found and the estimates of the true error?
- How do you think the results might change if there were more training points available?