**Report for Homework 2**

**1.Introduction**

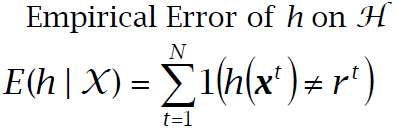
For a given training set for Class C, a most general hypothesis G and a most specific hypothesis S can be found to be used as a decision boundary in supervise learning. Also between G and S ,there exists a version space. Without knowledge of C , in order to find S and G, we can use the information of the training data to get the limits in X and Y for the location of S. After selecting out S , then we can extend the scope in X and Y to meet the limits for G. With S and G, we can also choose another hypothesis M which is in the middle of S and G. In terms of empirical error and generation capability, we can use K-fold cross validation to cut the data set N into training set and validation set to compute generation error for G,M and S. With the number N of data points increasing, we can get a set of generation errors for G,M and S. Therefore, we can figure out how the number of data points impacts S and G in terms of empirical error and generation capability.

**2.Methods**

In this work, 5 fold cross validation is applied. The whole data set N is cut into five parts. We randomly pick up one part as the validation set and leave the others as the training set. So for every number N of data points, we run the algorithm five times for different validation sets to get the average generation error for G,M and S .

In order to find out the location of S, the data points in training set are grouped into two matrix based on '0' and '1'. Then, the minimum and maximum value of X and Y can be found out in the data points of '0'. Therefore, we can locate the peaks for S based on these minimum and maximum values Xsmin, Xsmax, Ysmin, Ysmax. In this program, we create a function named bound to get the boundary for S. For this function, the input is the matrix for points with '1' , the output is Xsmin, Xsmax, Ysmin, Ysmax for S. After locating S , we cut the whole quadrant into 8 areas. Then by decreasing the value in X from S, we can search for the maximum value in X from the data points who satisfy X < Xsmin and Ysmin < Y < Ysmax. In the same way we can find out the minimum value in X from the data points who satisfy X > Xsmax and Ysmin < Y < Ysmax by increasing the value in X from S. Therefore, we find out the minimum and maximum value in X for G. Then based on Xgmin and Xgmax for G. The minimum and maximum Ygmin and Ygmax for G can be selected out in the same way from data points between Xgmin and Xgmax. However, in this way , we may find out two G because of the location of some specific points which can limit the extension of X or Y ,which may lead to different boundaries when starting from X and Y. In this program , we choose G which starts from the extension in X axis.

For every chosen validation set, we will compute the empirical error for the average generation capability for S ,M and G based on the formula shown below.



**3. Results**

In this program, as shown below, the green 'o' and blue 'plus' represent the data points in training set. And the black 'o' and 'plus' represent the data points in validation set. The blue rectangle represents Class C. And the yellow rectangle represents the most general hypothesis G. The magenta rectangle represents the hypothesis M between S and G ,and the red rectangle represents the most specific rectangle S. In fig.1 below, for training set , S and G have no empirical error, however for validation set , there is one black 'plus' outside S , so there is one empirical error for G which is recorded for the average error.

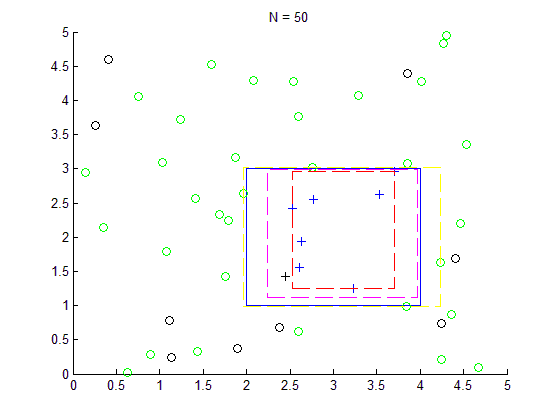


Fig.1 G,M,S when N = 50

We also draw the graph for N =100 ,N =250 and N =500. With the increasing number N, we can find G,S,M and C get closer and there are fewer empirical errors for G,M and S. In Fig. 4, G ,M and S almost become the same as Class C.

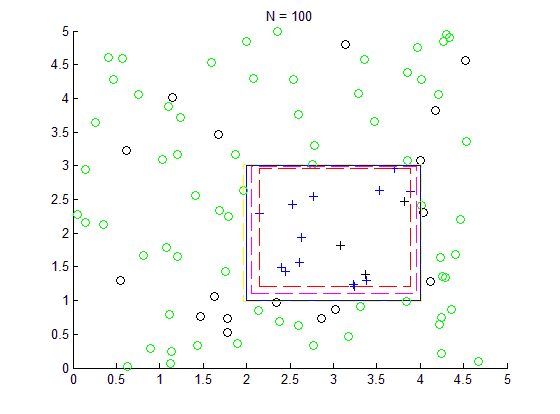


Fig.2 G,M,S when N = 100

In Fig.5 below, we compute the average generalization error with the increasing number N for G,S and M. We can find that when the number N is small such as N =50 , S,M and G may have different generalization errors. However , with the increasing of N , the errors will decrease significantly , also three errors for S,G and M trend to become almost the same which proves S,G and M are getting identical.

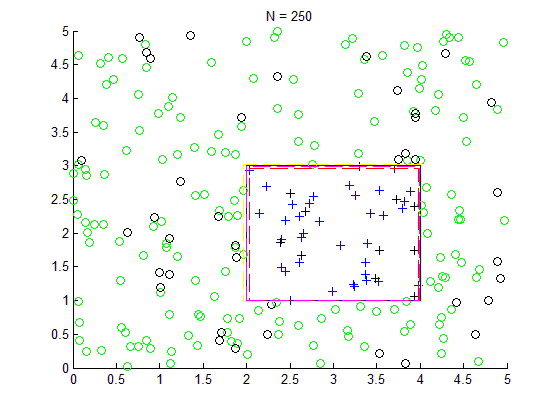


Fig.3 G,M,S when N = 250

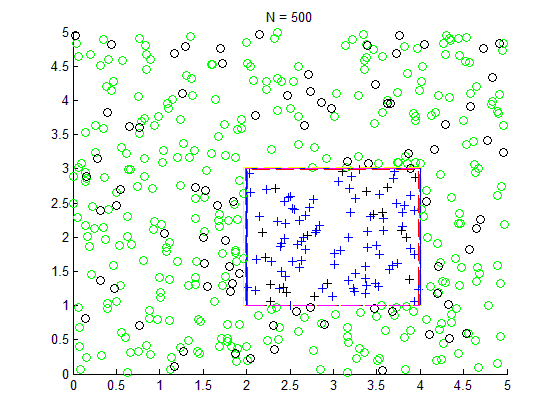


Fig.4 G,M,S when N = 500

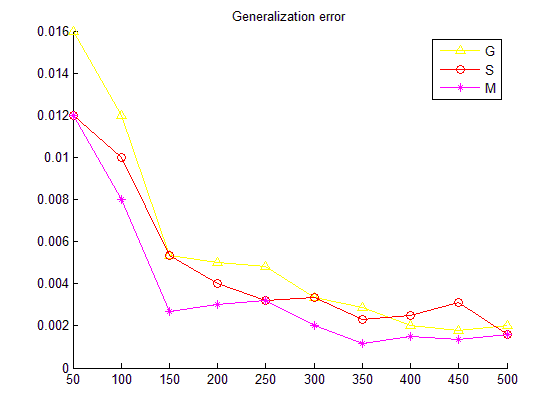


Fig.5 Average Generalization Error for G,S and M

**4. Discussion**

Based on the above results , we can find the average generalization error for S,M and G decreases significantly with the increasing of data set number N . Because we use 5-fold cross validation in this program, so the number of training set is about 0.8 of the data set number N. This means with the increasing of data set number N, the size of training set increases and the average generalization error for S,M,G decrease . When N =500, the generalization error becomes almost the same for S,M and G, which is about 0.002. Also we can find with the increasing of data set and training set, classifiers S,M and G trends to be identical as Class C.

**5. Software listing and executable software**

This program uses MATLAB including two files , HW2\_Lei\_Hu.m and bound.m . Open and run the file HW2\_Lei\_Hu .m to get the results and graphs.