Homework 4
Applied Machine Learning
Fall 2017
CSCI-P 556/INFO-I 526

Ashay Sawant

December 1, 2017

The work herein is solely mine.

Problem 1: K-Fold Cross Validation [25 points]

Implement k- fold cross validation and select $k = 5$ to create 5 training and 5 test data sets from each data set and save these 30 files. You will use these data sets for model comparison and parameter selection.
Answer:
I have divided the whole dataset, in five different folds based on the index of every row. From the five folds, I have created 5 different combinations for Training-Testing Dataset pair.

Problem 2: K-Nearest Neighbors (KNN)[55 points]

2.1 Implement KNN algorithm with two different distance functions. You can either use existing distance functions, i.e., Euclidean or design your own.
Answer:
KNN algorithm works well with numerical continuous data. Hence, I had to pre-process Car Evaluation Dataset and Credit Approval Dataset. As a Distance Metric, we have used Euclidean distance and Manhattan distance. Using these distances, we performed the normal voting to find the likelihood of testdata belong to a class.

2.2 Use the data sets obtained in problem 1 to determine the optimal k over each data set for KNN algorithm. For 5 different k values, plot the test error for each data set. Total number of figures = 3 (data set number) $\times$ 2 (distance function number) = 6. Report the best k and distance function for each data set.

Answer:

For this experiment, I have used 5 different values of K (1,9,19,29,49) to check how KNN performs on the various dataset.

- KNN over Ionosphere Dataset:

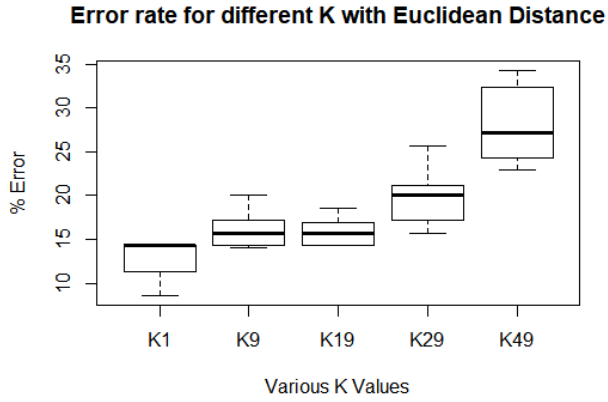

Figure 1: Error Rate in Ionosphere Dataset with Euclidean Distance

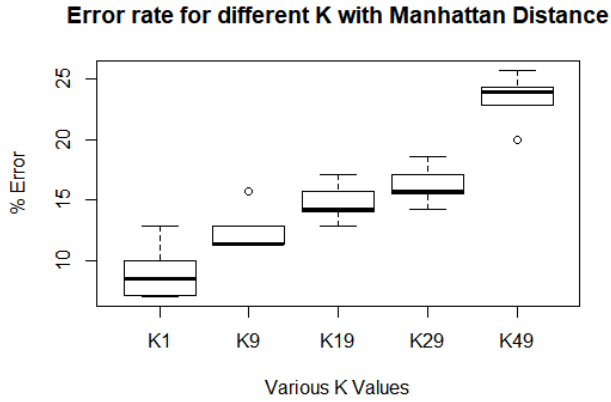

Figure 2: Error Rate in Ionosphere Dataset with Manhattan Distance

As we can see in boxplots, KNN performed almost same over Euclidean and Manhattan distance metric. Though, error rate was slightly low when we used Manhattan distance at K=1.

– KNN over Car Evaluation Dataset:

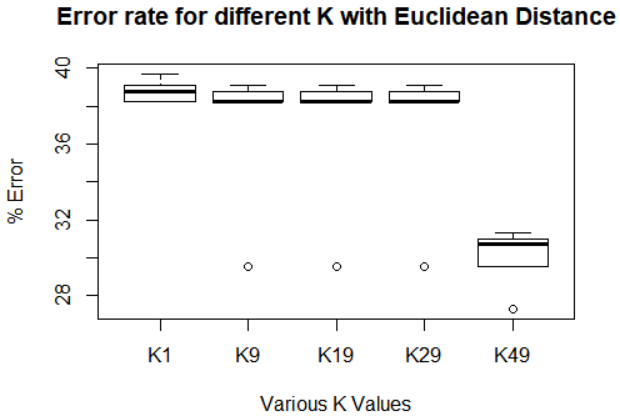

Figure 3: Error Rate in Car Evaluation Dataset with Euclidean Distance

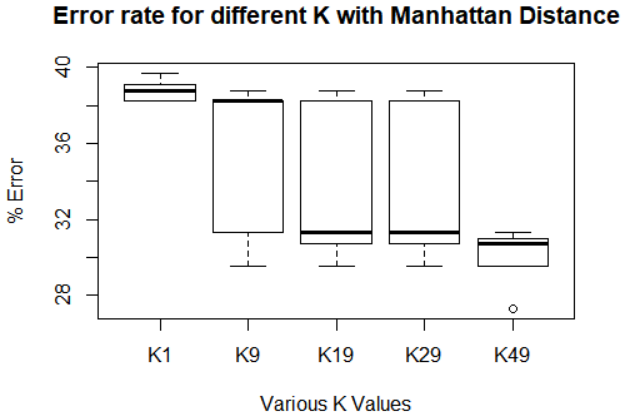

Figure 4: Error Rate in Car Evaluation Dataset with Manhattan Distance

As we can see in boxplots, error rate for KNN with Euclidean distance is saturated at fixed value whereas, with Manhattan distance it varied. So, we can conclude that with Manhattan Distance over Car Evaluation Dataset, KNN performed better at K=49.

– KNN over Credit Approval Dataset:

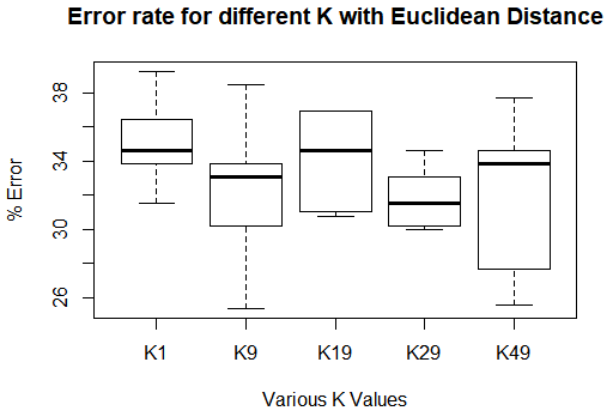

Figure 5: Error Rate in Credit Approval Dataset with Euclidean Distance

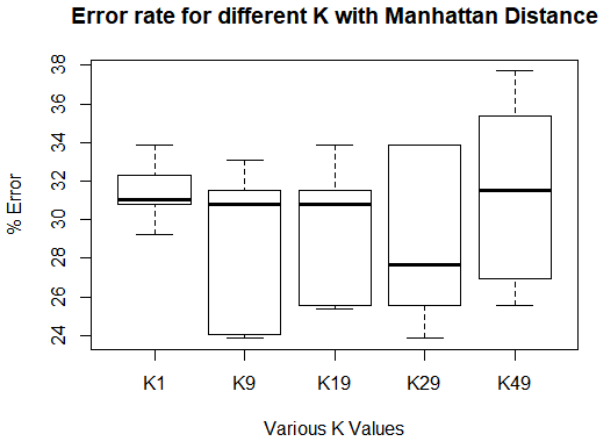

Figure 6: Error Rate in Credit Approval Dataset with Manhattan Distance

As we can see in boxplots, KNN performance was better with Manhattan Distance than Euclidean Distance. Lowest error rate counted was for Manhattan Distance with K=9.

2.3 Use the KNN package in R for validation.
Answer:

– KNN validation over Ionosphere Dataset:
As we can see in boxplots, error rate for KNN with R-package is minimum for K=1.

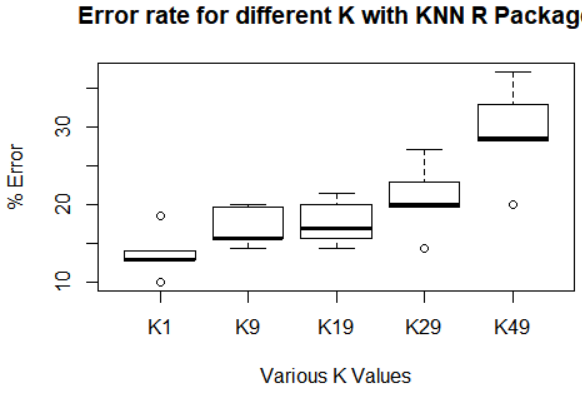

Figure 7: Error Rate in Ionosphere Dataset with R-package

– KNN validation over Car Evaluation Dataset:

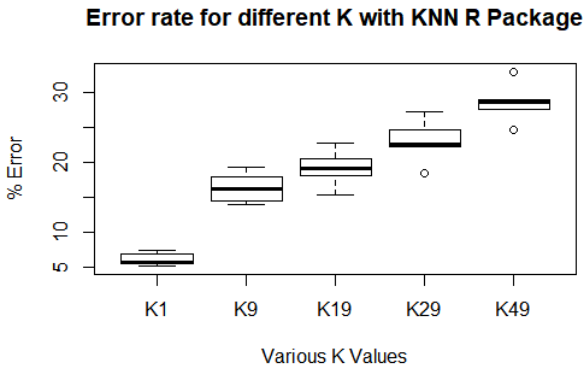

Figure 8: Error Rate in Car Evaluation Dataset with R-package

As we can see in boxplots, error rate for KNN with R-package is minimum for K=1.

– KNN over Credit Approval Dataset:

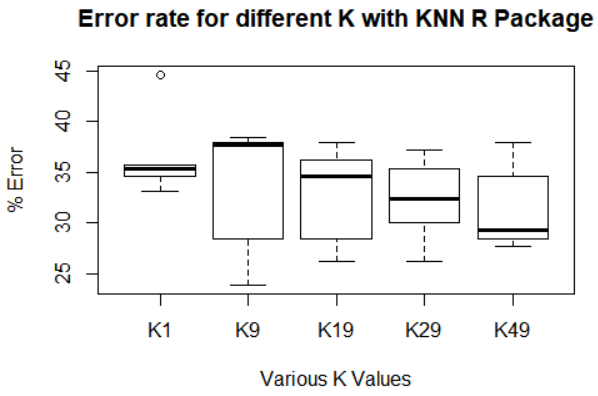

Figure 9: Error Rate in Credit Approval Dataset with R-package

As we can see in boxplots, error rate for KNN with R-package is minimum for both K=9 and K=19.

Problem 3: Naive Bayes Classifier [55 points]

3.1 Implement Naive Bayes classifier. The Pseudo-code for naive bayes algorithm is provided above. You may need to modify it for categorical variables. To handle unseen feature values, you may need to make use of m-estimate of conditional probability method. There are also other techniques, i.e., Laplace smoothing.

Answer:

The above problem requires us to develop a Naive Bayes Classifier for our given datasets. Naive Bayes algorithm is based on Bayes theorem. We have pre-processed the dataset, to convert continuous values into discrete values by binning them based to data distribution. To handle unseen feature values, I have used m-estimate of conditional probability method.

3.2 Train Naive Bayes classifiers over training data sets and test each classifier against corresponding test data. Make a plot that shows the error over each test data. Report the average error rate for 5-fold cross validation for each data sets.

Answer:

We have divided all three datasets into five mixtures for training-testing datasets randomly. As mentioned earlier, we have performed binning operation to factorize the dataset. Below graph shows the error rate on each fold in terms of mis-classified data.

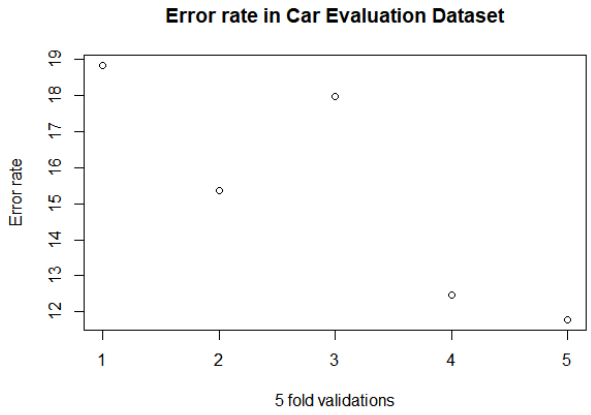

Figure 10: Error Rate in Car Evaluation Dataset

For Car Evaluation Dataset, the average error rate for 5-fold cross validation is 15.22%.

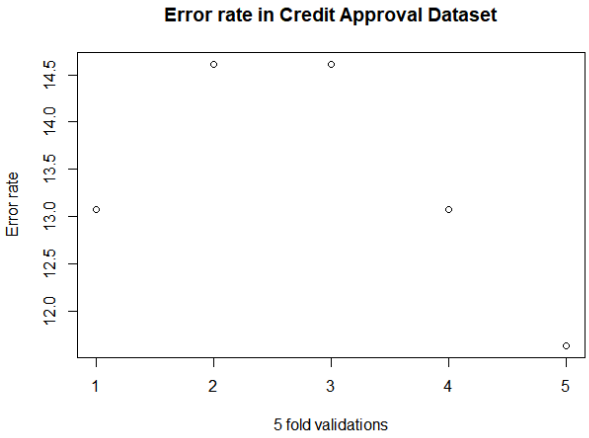

Figure 11: Error Rate in Credit Approval Dataset

For Credit Approval Dataset, the average error rate for 5-fold cross validation is 13.32%.

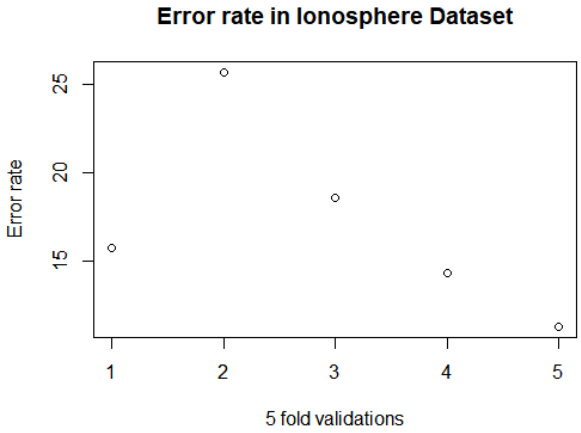

Figure 12: Error Rate in Ionosphere Dataset

For Ionosphere Dataset, the average error rate for 5-fold cross validation is 18.1%.

3.3 Use Naive Bayes package in R for validation.

Answer:

I have used **naivebayes** package for validation. Graph plots below, gives the Error rate in terms of % mis-classified data.

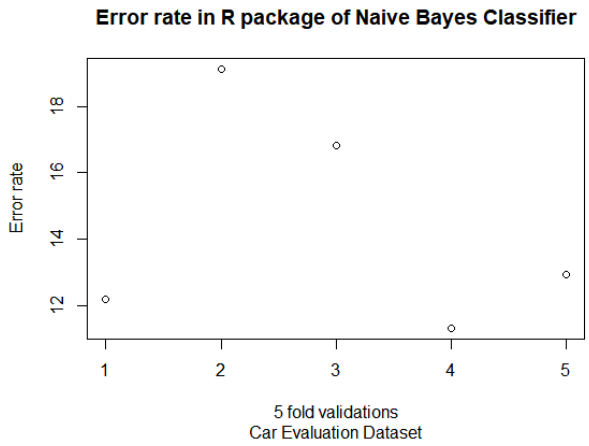

Figure 13: Error Rate in Car Evaluation Dataset

As we can see in plot above, the minimum error in Car Evaluation Dataset is between 11%-12%.

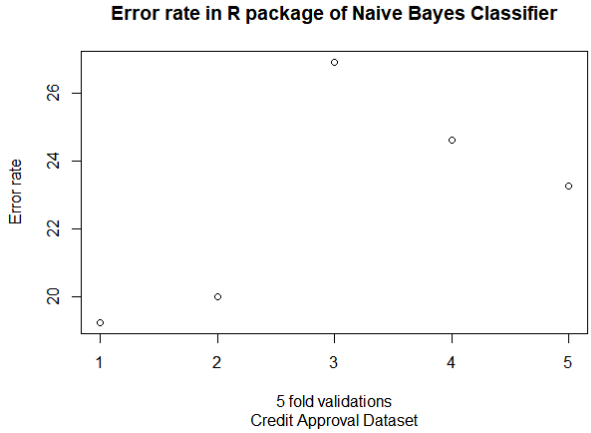

Figure 14: Error Rate in Credit Approval Dataset

As we can see in plot above, the minimum error in Credit Approval Dataset is between 19%-20%.

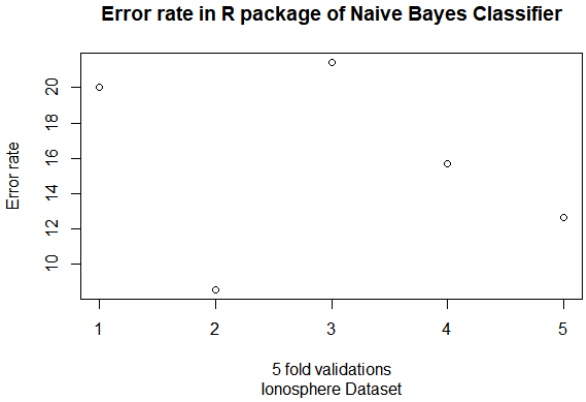

Figure 15: Error Rate in Ionosphere Dataset

As we can see in plot above, the minimum error in Ionosphere Dataset is between 9%-10%.

Problem 4: Naive Bayes Classifier vs. K-Nearest Neighbors [30 points]

In this question, you are asked to compare Naive Bayes classifier with k-nn algorithm. First, determine the best KNN model for each data set. Then, Make a plot that reveals comparison of two algorithms using test error for each data set. (Total number of figures = 3)
Answer:

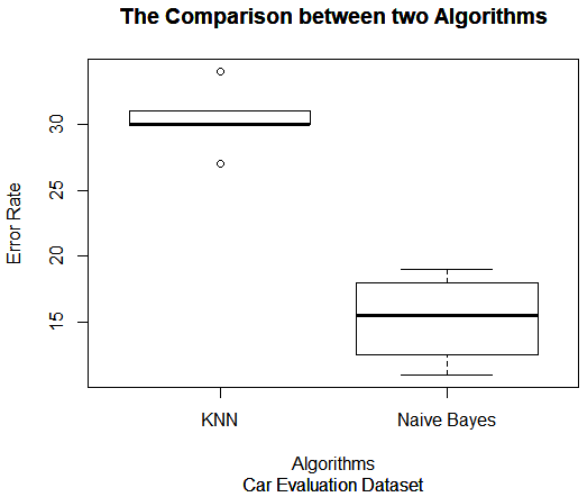

Figure 16: Comparison with Car Evaluation Dataset

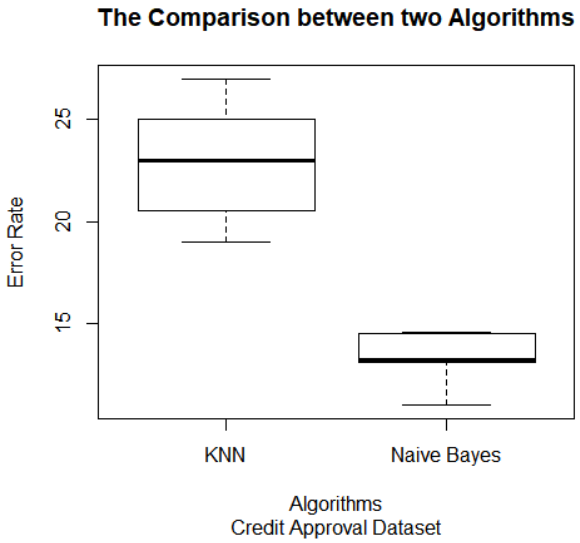

Figure 17: Comparison with Credit Approval Dataset

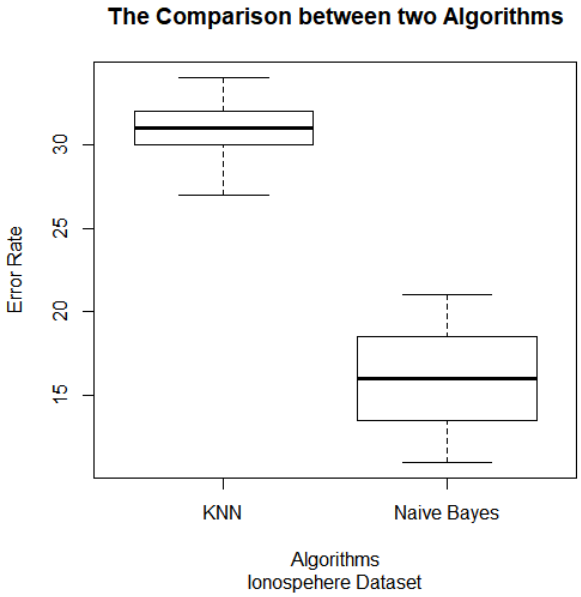

Figure 18: Comparison with Ionosphere Dataset

As we can see all the three box-plot for respective datasets, we can clearly say that Naive Bayes performed better in terms of Error Rate than KNN Algorithm.

Problem 5 [15 points]

From textbook, Chapter 4 exercise 10.g and 13 (only for k-nn and logistic regression)

Question 10.g - Answer:

In this question, we asked to fit k-nn model with K=1 on Weekly dataset with "Lag2" as the predictor. We have used knn from "class" package in R. The Error rate for trained model is 50.9615%. The confusion matrix for actual and predicted labels is as follows:

Prediction	Down	Up
Down	21	29
Up	22	32

Question 13 - Answer:

In this question, we are asked to fit logistic regression and k-nn over Boston dataset to predict whether a given suburb has a crime rate above or below the median. Error rate in Logistic Regression and KNN for various values of K is as follows:

Logistic Regression	18.18182
KNN(K=1)	45.8498
KNN(K=5)	16.99605
KNN(k=10)	11.46245

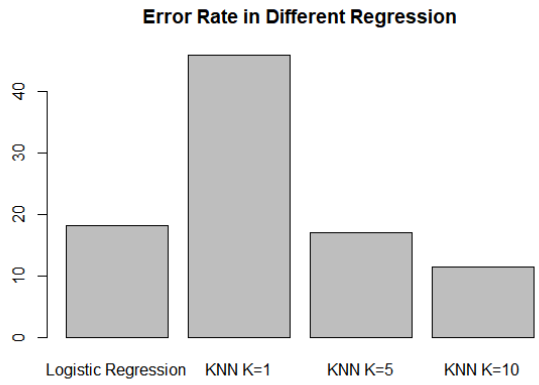

Figure 19: Comparison between Different Regression Problem

Extra credit (optional) [40 points]

1. You are asked to evaluate the performance of two classifier models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as A through Z. The table below shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(-) = 1 - P(+) and P(-|A,...,Z) = 1 - P(+|A,...,Z). Assume that we are mostly interested in detecting instances from the positive class.

   (a) Plot the ROC curve for both M_1 and M_2. Which model do you think is better. Explain your reasons?
   Answer:
   The following graph denotes the ROC curve for both M_1 and M_2:

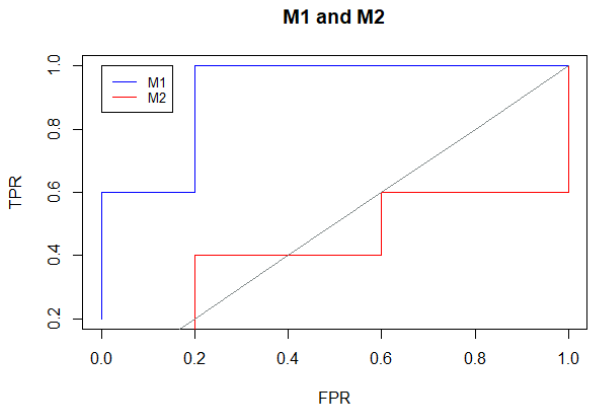

Figure 20: ROC curve for M_1 and M_2

M1 is better,because it has more area under the ROC curve than M2.

   (b) For model M_1, suppose you choose the cutoff threshold to be $t = 0.5$. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.
   Answer:
   For this problem we find out the confusion matrix for M_1 is shown below:

Actual	+	-
+	3	2
-	1	4

Thus, the precision can be calculated as $3/4 = 75\%$
Recall will be $3/5 = 60\%$
F-measure is $(2 * 0.75 * 6)/(0.75 + 0.6) = 0.667$

(c) Repeat the analysis for part (c) using the same cutoff threshold on model M_2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

Answer:

For this problem we find out the confusion matrix for M_2 is shown below:

Actual	+	-
+	1	4
-	1	4

Thus, the precision can be calculated as $1/2 = 50$ %

Recall will be $1/5 = 20$ % .

F-measure is $(2 * 0.5 * 0.2)/(0.5 + 0.2) = 0.2857$

From the above values,M_1 is still better than M_2. So, the result is consistent.

(d) Repeat part (c) for model M_1 using the threshold $t = 0.1$. Which threshold do you prefer, $t = 0.5$ or $t = 0.1$? Are the results consistent with what you expect from ROC curve?

Answer:

For this problem we find out the confusion matrix for M_2 is shown below:

Actual	+	-
+	5	0
-	4	1

Thus, the precision can be calculated as $5/9 = 55.6\%$

Recall will be $5/5 = 100\%$.

F-measure is $(2 * 0.556 * 1)/(0.556 + 1) = 0.715$

Even though, from the value of F-measure, t=0.1 is better than t=0.5, but since $(0.2, 0.6)$ which are FPR and TPR when t=0.5, is closer to $(0, 1)$ than $(0.8, 1)$, FPR and TPR when t=0.1, we prefer t=0.5 over t=0.1.

This inconsistency can be shown by computing the area under the ROC curve:

For t=0.5, area$=0.6 * (1 - 0.2) = 0.6 * 0.8 = 0.48$.

For t=0.1, area$=1 * (1 - 0.8) = 1 * 0.2 = 0.2$.

From the above values, we can see that area for t=0.5 is greater than t=0.1 Thus, we prefer t=0.5