Mohammad Esfandiari

Assignment 1

1) Evaluating a classification algorithm

a) As you probably know in order to measure the performance of a classification algorithm, one can simply find the number of misclassified points over all the points (E). A better version of the previous method is Balanced Error Rate where you divide the error by the size of each class (BER). Another important measure in the world of computer science and statistics is Precision/Recall. Precision is the probability of an assignment of true positive labels over all the positive labels. Recall is the probability of an assignment of true positive labels over all the predicted labels. Using the following table (Table 1) we will explore these measures:

Output Labels	True Labels		
		1	-1 (0)
	1	True Positive (TP)	False Positive (FP)
	-1(0)	False Negative (FN)	True Negative (TN)

Using the above table we can easily rewrite the above measurements as:

$$\boldsymbol{E = \frac{FP + FN}{TP + FP + FN + TN}}$$

$$BER = \frac{1}{2} \times \left(\frac{FP}{FP + TN} + \frac{FN}{FN + TP} \right)$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

b) Given the Following data about True Label and Predicted Labels, compute E, BER, Precision and Recall. Show your steps.

Data Point	True Labels	Predicted Labels
0	1	1
1	1	-1
2	1	-1
3	1	-1
4	-1	1
5	-1	1
6	-1	1
7	-1	-1
8	-1	1
9	-1	-1
10	-1	-1
11	1	1
12	-1	1
13	-1	-1

Table 1

c) Write a program in Python that given true labels and predicted labels compute the following:

TP, FP, TN, FN, E, BER, Precision, Recall.
Your program should follow these rules:

i) It must read the data from a file that the user will provide. Hard coded filenames are not acceptable.

ii) It must print out the all the information mentioned above in separate lines.

iii) The code must be commented and well organized. Also be careful about the tabs and spaces in Python.

iv) Your code must be compatible with Python3.x (i.e version 3 of Python)

v) Follow the naming conventions mentioned before.

vi) Be sure to run your codes on OSL machines.

d) Can you form an example in which each one of these measures does not fully reflect the power of classification method? Which one do you think works better in general?

2) Nearest means Vs. Naïve Bayes

a) Nearest Means and Naïve Bayes are two basic classifiers in Machine Learning. The idea behind these classifiers is simple. Given a set of points which corresponds to multiple classes find the average of points in each class and for a given new point just compare the distances to the means. The formal definition of Nearest Means for two classes is as follows :

i) Given Training Set $T \subseteq \mathbb{R}^{N \times M}$ and set of labels L where $l_i \in \{-1, 1\}$ and a Test Set $E \subseteq \mathbb{R}^{N \times M}$

ii) Training Step:
For each l_i compute $m_i = \text{ mean of } t_j \in T \; where \; L(t_i) = l_i$

iii) Prediction Step:
For every $x_i \in E : L(x_i) = \min_j \left\| m_j - x_i \right\| \; where \; j \in \{-1,1\}$

b) Naïve Bayes follows the same principle but with two major difference. Below is the steps of a Naïve Bayes algorithm :

i) The same as Nearest Means.

ii) Training Step:
For each l_i compute $m_i = \text{ mean of } \forall \text{ j } t_j \in T \; where \; L(t_j) = l_i$
For each l_i compute $s_i = \text{ variance of } \forall \text{ j } t_j \in T \; where \; L(t_j) = l_i$

iii) Prediction Step:
For every

$$x_i \in E : L(x_i) = \min_j \sum_{k=1}^{m} \left(\frac{x_{ik} - m_{jk}}{s_{jk}}\right)^2 \; where \; j \in \{-1,1\}$$

c) Implement Nearest Means algorithm for more than two classes. Basically you need to write a program that gets the data file and training labels from the user and output the means of all the classes in the standard output and also write the prediction labels in a file named $\{\text{name\_of\_datafile}\}.\text{prediction}$
d) Implement Naïve Bayes algorithm for more than two classes. Basically you need to write a program that gets the data file and training labels from the user and output the means and the standard deviation of all the classes in the standard output and also write the prediction labels in a file named $\{\text{name\_of\_datafile}\}.\text{prediction}$
e) Run both algorithm on Breast Cancer, Qsar, Inosphere, Climate Change dataset and conclude which one has higher accuracy.
f) What happens when the variance of the data is zero ? How can we change the algorithm to get rid of this problem?