Miniproject 1
BMI 555 IEE 520
Fall 2020

Due Date: September 15, 2020

1) Use a Naïve Bayes classifier. Complete the following calculations **without** a Naïve Bayes software package so you understand the steps. You might use Microsoft Excel.

a) Fill the predicted class for each instance and record your prediction in the right-most column in the table.

b) Create the confusion matrix for all three classes.

c) Assume that class 2 is the positive class. Calculate TPR/FPR and plot a ROC curve for this model and find the area under the curve.

Instance	X_1	X_2	X_3	Y	Predicted Class ($\hat{Y}$)
1	4	Sunny	High	2	
2	3.7	Rainy	Normal	2	
3	12	Rainy	High	1	
4	10	Sunny	Normal	0	
5	24	Rainy	Normal	2	
6	28	Rainy	High	2	
7	18.4	Sunny	Normal	0	
8	7.2	Sunny	Normal	1	
9	36	Rainy	High	1	
10	34	Rainy	Normal	2	

Hints:

- In order to find the predicted class ($\hat{y}$), we use the Bayes formula as follows:

  $P(Y = y|\, X = x) = \frac{P(X=x|\, Y=y).P(Y=y)}{P(X=x)}$

  Because the values in the denominators are equal for all classes, we only compare numerators with each other, and then the predicted class label will be found by finding which class has the most value in the posterior value. But we can also compute the denominator from Bayes formula to determine the probability $P(Y = y|\, X = x)$ for each instance and class.

  Also, because we have conditional interdependencies, the formulae can be rewritten as follows:

  $P(X = x|\, Y = y)P(Y = y) = P(X = x_1|\, Y = y)P(X = x_2|\, Y = y)P(X = x_3|\, Y = y)P(Y = y)$

- Note that since X_1 is continuous, we need to estimate the underlying probability distribution function to calculate $P(X = x_1|\, Y = y)$.
- The prior probabilities for actual labels can be calculated with $P(Y = y) = \frac{N(Y=y)}{N}$ where *N* is the function which calculates the number of instances which satisfy the condition.

2) Build a **Naïve Bayes classifier** for the given data set in Python. Provide the code and the results of your analysis.

a) Evaluate the generalization error of the classifier in three ways:

- From the training data
- From a test set of 20% of the data
- From cross-validation with 5 folds

b) Comment on any differences in these estimates of generalization error.

c) Provide the code and a confusion matrix, summary statistics, and ROC curves calculated from the **cross-validation** only.

About data set:

The Avila data set has been extracted from 800 images of the 'Avila Bible', an XII century giant Latin copy of the Bible. The prediction task consists in associating each pattern to a copyist[1].

Attributes Information:

- F1: intercolumnar distance
- F2: upper margin
- F3: lower margin
- F4: exploitation
- F5: row number
- F6: modular ratio
- F7: interlinear spacing
- F8: weight
- F9: peak number
- F10: modular ratio/ interlinear spacing
- Target Value Classes: A, B, C, D, E, F, G, H, I, W, X, Y

[1] https://archive.ics.uci.edu/ml/datasets/Avila