SEng 474 / CSc 578D
Data Mining – Fall 2016
Assignment 2

Due: October 14th – NO LATE ASSIGNMENTS ACCEPTED.

Be aware: no late assignments will be accepted so we can release the key on the due date to help you prepare for the midterm.

Submit all answers through ConneX before 11:55pm.
Different marking schemes will be used for undergrad (SEng 474) and grad (CSc 578D) students. Undergrad students do not have to answer the grad questions.
Always show your work. **No marks will be given for answers only.**

All code questions are implemented in python 2.7.

1. Classifier Accuracy (SEng 474 and CSc 578D: 10 points)
Assume you were given a dataset built from random data, where attributes values have been randomly generated with no consideration to the class labels. The dataset has three classes: "red", "blue" and "yellow". You were asked to build a classifier for this dataset, and told that 50% of the data will be used for training, and 50% for testing. The testing set is balanced, so you can assume it has the same distribution as the training set. Because you are smart, you will start by establishing a theoretical baseline for your classifier's performance.

a) (2 points)
Assume the data is equally split between the three classes (33.3% "red", 33.3% "blue" and 33.3% "yellow") and your classifier systematically predicts "red" for every test instances, what is the expected error rate of your classifier? (Show your work)

b) (3 points)
What if instead of always predicting "red", the classifier predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3. What is the expected error rate of the classifier in this case? (Show your work)

c) (2 points)
Now lets assume that the data is not split equally, but has half (1/2) of its data labeled "red", one-fourth (1/4) labeled as "blue", and one-forth (1/4) labeled as "yellow". What is the expected error rate of the classifier if, as in question a), the prediction is "red" for every test instances.

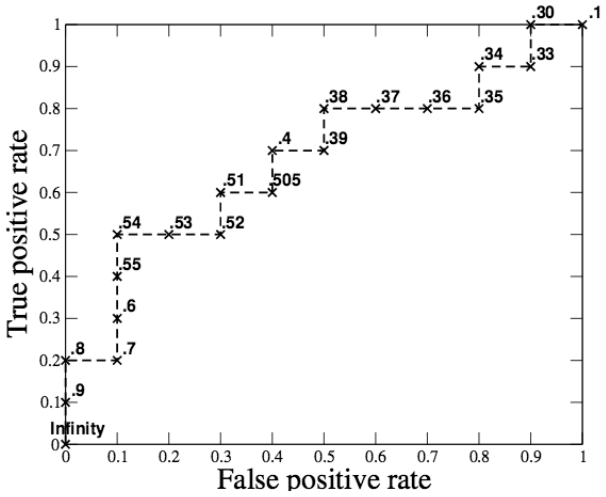
d) (3 points)
With this dataset (three-fourth labeled "yes", one-fourth labeled as "no) What is the expected error rate of the classifier if, as in question b), it predicted "red" with a probability of 0.7, and "blue" with a probability of 0.3. (Show your work)

Receiver Operating Characteristics (ROC) graphs
You have been introduced to different measures of performance for classifiers: accuracy, precision, recall, F-measure etc. It is important to look at them all, as one measure can hide important information. For example, a 60% accuracy obtained by a classifier dealing with 20 classes (labels) is doing much better than one obtaining a 60% accuracy on 2 classes.
ROC graphs are two-dimensional graphs plotting the *True Positive rate (TP rate)* on the y-axis, against the *False Positive rate (FP rate)* on the x-axis (See the paper in [1] for definitions). A discrete classifier produces TP and FP rates, represented by a single point in the ROC graph.

The point at (0,0) represents a classifier that always predicts negative. This strategy assures that the classifier never generates a false positive. However, this strategy would also never generate a true positive. The upper left corner (0,1) represents a classifier that yields a perfect classification. Finally, the upper right corner (1,1) represents a classifier that would always issue a positive classification, generating true positives and false positives equal to the total number of positive and negative instances in the test set (respectively). Thus, the TP rate and FP rate are both 1.

Classifiers such as Naïve Bayes and Neural Networks typically issue an instance *probability* or *score* along with each of their predictions. For example, the Naïve Bayes classification question on your last assignment chose the class with the highest probability. Imagine that instead of choosing the class with the highest probability, we choose a class when its probability is greater than some threshold "t". ROC graphs are often used to visualize the performance of such classifiers at different score thresholds.

Given a list of instances, their classes and their scores, we can plot the ROC graph. Scores are plotted one by one **in descending order**. Each instance is a step moving up if positive, and right if negative. The length of the steps going up depends on the total number of positive instances, and the length of the steps moving right depends on the number of negative instances. For example, if the list has 10 positive instances, it will move up by 0.1 each time. If it has 5 negative instances, it will move right by 0.2. Figure 3 is given as an example of a ROC graph built to analyze the performance of a classifier at different threshold.



Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Fig. 1. From the paper *ROC graphs: Notes and practical considerations for researchers*.[1]

Looking at Figure 1, we can see that if we chose to use a classifier with a threshold of 0.8, we get a *precision* of 100% (2/2), but a *recall* of 20% (2/10). With a threshold of 0.54, we lost in *precision* with 83.3% (5/6), but the *recall* has increased to 50% (5/10).

2. ROC Curves (SEng 474: 20 points; CSc 578D: 30 points)

You are asked to evaluate the performance of two classifiers, A and B. The following table shows the ranking obtained by applying the classifiers to a test set of 10 instances.

Instance	True Class	Classifier A	Classifier B
1	P	0.73	0.61
2	P	0.69	0.03
3	N	0.44	0.68
4	P	0.55	0.31
5	N	0.67	0.45
6	N	0.47	0.09
7	P	0.08	0.38
8	N	0.15	0.05
9	N	0.45	0.01
10	P	0.35	0.04

a) (10 points) Plot the ROC graphs for both A and B on the same graph, as in Fig 1.

b) (2 points) For classifier A, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instances whose ranking is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the classifier at this threshold value.

c) (2 points) Repeat the analysis for part (b) using the same cutoff threshold on classifier B. Compare the F-measure results for both classifiers. Which classifier is better? Are the results consistent with what you expect from the ROC curve?

d) (3 points) Plot the curve of an unbiased random classifier (equal probability of predicting positive or negative) on the graph in **a)**. At what threshold does classifier A performs better than a random classifier? At what threshold does classifier B perform better than a random classifier?

e) (3 points – both SEng 474 and CSC 578D):
Plot the curve of a random classifier that predicts positive with a probability of 0.7, and negative with a probability of 0.3.

f) For grad students (CSc 578D: 10 points):
Based on the Tom Fawcett's paper [1]: ROC graphs: *Notes and practical considerations for researchers*:

I) (CSc 578D: 5 points) Consider two discrete classifiers whose performances have been placed on ROC graph. Classifier A's coordinate are (0.3, 0.7) and classifier B is positioned at (0.8, 0.1). Which of the two classifiers would you choose and why?

II) (CSc 578D: 5 points) Briefly explain how we could create a ROC graph to evaluate the classifiers from question 1 ("red", "blue", "yellow").

3: MLE and MAP estimates (SEng 474: 10 points; CSc 578D: 10 points)

a) Let θ = P(X=T). Calculate the MLE for θ for the following dataset by finding the maximum of P(D| θ). [4 marks]

D={T, T, T, T, T, T, T, F, F, F}

b) Recall the PDF for a Beta random variable is proportional to

$$\Theta^{(\beta 1-1)} * (1-\theta)^{(\beta 2-1)}$$

with parameters β1 and β2. Let's say you have evidence from previous studies that P(X=T) = ½. Let β1 = 4. Find β2 and then calculate the MAP estimate P(θ | D) for θ with a Beta(β1, β2) prior and the dataset above. [6 marks]

4: Gradient Descent (SEng 474: 20 points; CSc 578D: 20 points)

We have a new dataset where the input data has 2 continuous variables (x_1 and x_2), and the task is to predict a continuous value as the output y (i.e. regression). We have reason to believe the following new model for regression is a better fit to the problem domain:

$$\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2}^2$$

A) Write down the error function for this model. You should use the sum of squared error, as in class: [5 marks]

$$E(X) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

B) Derive the gradient descent update for w_0 using learning rate κ [5 marks]

C) Derive the gradient descent update for w_1 using learning rate κ [5 marks]

D) Derive the gradient descent update for w_2 using learning rate κ [5 marks]

References

1. Fawcett, Tom. "ROC graphs: Notes and practical considerations for researchers." *Machine learning* 31 (2004): 1-38.