CS 5525: Data Analytics I

Homework 3

Due Date: October 18th, 2016 (4:00PM) **Total: 100 Points**

Problem 1. Evaluation Measures

(10 Points)

For the Confusion Matrix shown below, compute the following values:

Confusion Matrix	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	100	25
	-	50	145

(a) Accuracy (b) Precision (c) F-measure (d) Sensitivity (e) Specificity (f) Cost (classification cost is -1 and misclassification cost is 5).

Problem 2. Precision-Recall

(15 Points; 6 + 9)

A database contains 100 documents out of which only 10 documents are relevant for a given query. Two search engines, A and B, report the following documents.

System A: **RRNNNRNRNN RNNRNNRNRN NNRNNNNNRN** and so on
System B: **RNRNRNNNNN NRRNNNNNNR NNNNNRNNNN** and so on

where **R** represents relevant document and **N** represents Non-relevant document. A search engine expert wants to estimate the accuracy of both these systems by computing the following information, so that, he can know how many documents to display.

(a) Using the PR-Curve info, compute the Precision at 40% Recall.
(b) If the expert decides to display only the first 15 documents for both of these search engines, calculate Precision, Recall and F-measure.

Problem 3. ROC Values

(20 Points; 10 + 10)

PART I: Evaluate the performance of two classification models M1 and M2 for a two class problem. The test set consists of a set of attributes (A through Z). The following Table shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are given).

Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

(a) Compute the Recall and F-measure for the model at this threshold value of 0.5 for both models M1 and M2. Show the corresponding contingency tables.
(b) For both of these models, what is the precision when the recall is 60%.

PART II: You have been asked to develop a classification model for diagnosing whether a patient is infected with a certain disease. To help you construct the models, your collaborator has provided you with a small training set (N = 10) with equal number of positive and negative examples. You tried several approaches and found two most promising models, C1 and C2. The outputs of the models in terms of predicting whether each of the training examples belong to the "positive" class are summarized in the table below. The first row shows the probability a training example belongs to the positive class according to classifier C1, while the second row shows the same information for classifier C2. The last row indicates the true class label of the 10 training examples.

$P(y=+\|C_1)$	0.1	0.15	0.2	0.3	0.31	0.4	0.62	0.77	0.81	0.95
$P(y=+\|C_2)$	0.25	0.49	0.05	0.35	0.66	0.6	0.7	0.65	0.55	0.99
y	-	+	-	-	+	-	+	+	-	+

(a) Draw the ROC curve for both the classifiers. (NOTE: You will just need to plot the points of the true positive rate at different false positive rates marked on the x-axis. Finally, join those points with line). Which classifier has a larger area under the ROC curve? (HINT: You need to consider different recall rates and compute the corresponding true positive and false positive rates)

(b) Wilcoxon Mann Whitney statistic
Compute the Wilcoxon Mann Whitney statistic for both classifiers. The statistic can be computed as follows:

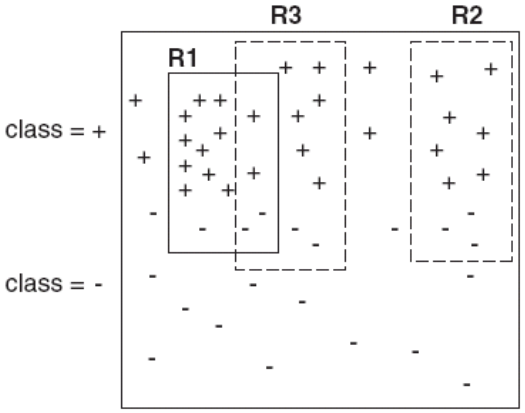
$$\text{WMW} = \frac{\sum_{i=0}^{m-1}\sum_{i=0}^{n-1} I(x_i, y_i)}{mn}$$

where $I(x_i, y_i) = 1$ *if* $x_i > y_i$ *and 0 otherwise.* Note that $\{x_0, x_1, \ldots, x_{m-1}\}$ correspond to the classifier outputs for the ***m*** positive examples while $\{y_0, y_1, \ldots, y_{n-1}\}$ correspond to the classifier outputs for the ***n*** negative examples (in this exercise, $m = n = 5$). Which classifier has a larger WMW value?

Problem 4. Rule-based Classification **(30 Points; 15+15)**

PART I: The following Figure illustrates the coverage of the classification rules R1, R2, and R3. Determine which is the best and the worst rule according to:

(a) FOIL's Information Gain.
(b) The m-estimate measure (with $k = 2$ and $p+ = 0.58$).
(c) The rule accuracy after $R1$ has been discovered, where only the positive examples covered by $R1$ are discarded.

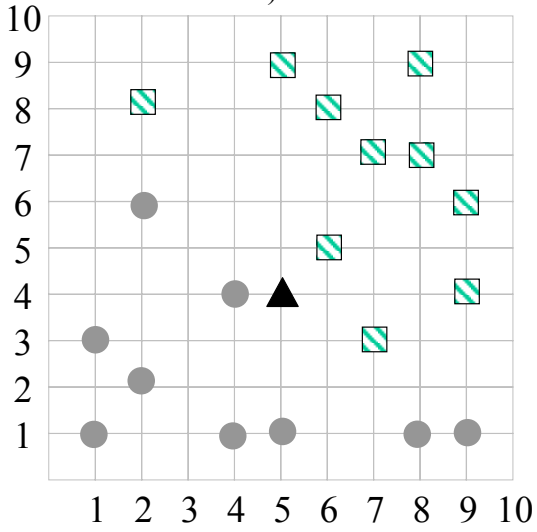


PART II: Consider a training set that contains 100 positive examples and 200 negative examples. For each of the following candidate rules compute the coverage, accuracy and FOIL's Information Gain. (Assume that the initial rule $R0$ is $\varnothing \rightarrow +$)

R1 covers 12 positive and 3 negative examples,
R2 covers 20 positive and 10 negative examples,
R3 covers100 positive and 80 negative examples.

Problem 5. Nearest-Neighbor Classification **(25 Points; 10+15)**

PART I: Consider the following 2-dimensional dataset. Classify the test point (triangle) using: (Treat the squares as + and circles as –)



(a) 5- nearest neighbor
(b) Manhattan distance weighted 3-nearest neighbor (the weight is $1/d^2$)

PART II:

Consider the three-dimensional data set in train.csv.

(a) Classify the data points in test.csv according to their 3-nearest neighbors. Also, give the probability estimates for the final decision.
(b) Do the same for the Euclidean distance weighted 3-nearest neighbors ($1/d^2$). Does the predicted label for each point remain the same as that in (a)?
(c) In the test.csv, the true class labels are also provided. Construct the confusion matrix and calculate Accuracy, Precision, F-measure for problems (a) and (b). From your results, which method gives better performance?

Please also submit your source code for full credit.