CS5567 Spring 2016 Assignment I. (70+30) pts.

NAME(s) (use Pathway format):

Raghunandan Rao Malangully: rm9vf

Santhosh Kumar Gattu: sg6n6

Hirenbhai Harshadbhai Shah: hhstm4

Rakesh Reddy Bandi: rb8xf

Word processed electronic submission on Blackboard is due latest by 11 pm on Wednesday, Feb 10th. Submissions received after the deadline will be graded only for effort for a maximum of 70% of the total grade (Refer to class syllabus for detailed grading policy). State any assumptions you make, justify your answers, show intermediate steps and explain your results for maximum credit. For machine learning questions, the answer should be written in the form of a concise scientific report with a focus on rationale and evaluation. The report should be a polished depiction of what you did, why you did it, how exactly you did it, and how well it worked.

All answers should be in your own words with any sources you refer to cited at the appropriate places. Any knowledge you acquire from the Internet should be written in your own words and be appropriately referenced. Copying and pasting from the Internet, each other or any other source will not count as your effort (Refer to class syllabus for detailed policy on plagiarism).

You may submit this assignment in groups of up to 4 each. Write your names on this sheet and include it as the cover page for your submission. The submission should consist of a standalone word-processed report file and additional files (code, data, instructions to run your code). Code should be submitted as a text file (not copied and pasted into Word) annotated with comments (include your names within the code files).

Compress all files into an archive before uploading on Blackboard. Be sure to name your archive file using the names of group members, e.g., HW1\_RadhaKrishnaRomeoJuliet (and not just “HW1”). If you upload the wrong file, don’t panic. You may upload upto 3 times.

You may use any programming language or package of your choice for programming assignments in the homework. Possible choices are R, MATLAB, Python (e.g., sci-kit or code on textbook website) or JAVA (e.g., Weka). Be sure to cite your source if you use pre-existing code. Clearly distinguish between pre-existing code and your contributions.

Q1. (30) Measures like precision, recall and specificity have the general form (x/x+y) where x and y are drawn from the set {TP, TN, FP, FN}. List all possible expressions of this form and evaluate them as possible measures of performance. For each measure, comment on whether it captures a meaningful aspect of predictive performance. For each meaningful measure, provide two real world scenarios, one where the measure would be useful and one where the measure would be a bad choice.

Q2. (60) The goal of this assignment is to use the Naïve Bayes Classifier to predict if a tennis player won or lost a match based on his/her match statistics. Relevant data is available on the US (<http://www.usopen.org/index.html>) and Australian (<http://www.ausopen.com/index.html>) Open websites. You may download data manually or write a script. Make appropriate corrections for differing length of matches. Use an appropriate scheme for objective evaluation and reporting predictive accuracy.

Q3. (10) Identify and correct at least one mathematical typo in Chapter 2.

**Solutions**

1. **Measures like precision, recall and specificity have the general form (x/x+y) where x and y are drawn from the set {TP, TN, FP, FN}. List all possible expressions of this form and evaluate them as possible measures of performance. For each measure, comment on whether it captures a meaningful aspect of predictive performance. For each meaningful measure, provide two real world scenarios, one where the measure would be useful and one where the measure would be a bad choice.**

**Solution**.

Following are the possible expressions:

1. Accuracy: It is the statistical measure that defines the proportion of total number of predictions that were correct. By using Accuracy, we can measure how often the classifier is correct.

Accuracy = *TP* + *TN / (TP* + *FP* + *TN* + *FN)*

In case of meteorological measurements Accuracy is a good choice.

In case of cricket bowling machine Accuracy is a bad choice

1. Sensitivity: It is the statistical measure of proportion of number of positives identified. It is also called as Recall / True positive rate. It defines the Positive predictive value

Sensitivity = TP / (TP+FN)

In case of categorizing a person having disease measuring Sensitivity is a good choice.

In case of a categorizing a disease free person measuring Sensitivity is a bad choice.

1. Specificity = It is the statistical measure of proportion of number of negatives identified. It is also called as True Negative rate. It defines the Negative predictive value.

Specificity = TN / (TN+FP)

In case of categorizing a disease free person measuring Specificity is a good choice.

In case of a categorizing a person with disease measuring Specificity is a bad choice.

1. Precision: It is the fraction of retrieved instances that are relevant. It is also called as Positive predictive value (PPV).

Precision: TP / (TP +FP)

In case of checking whether a mail is spam or not is a good choice for Precision. We can optimize precision to minimize false positives, where non spam has predicted as spam.

In case of testing whether a player throws a ball into the basket is a bad choice for Precision.

Precision alone can’t be considered as helpful measurement in terms of predictive analysis. We need to look at accuracy as well.

1. Negative Predictive Value (NPV): It is the fraction of retrieved instances that are irrelevant.

NPV = TN / (TN+FN)

In case of a disease, doctor is predicting that a patient doesn’t have disease is a good example.

Where doctor is predicting that a patient has disease is a bad example.

1. False Positive Rate (FPR): It is the proportion of absent events that yield positive test outcomes. It is also referred as Fallout.

FPR = FP/ (FP + TN) = (1- Specificity)

Consider the scenario of a health checkup camp where everyone is tested for a disease, irrespective of their health status. FPR would be a good measure in such scenario.

The scenario where disease affected people have to be vaccinated, this measure would be inappropriate one.

1. False Negative Rate (FNR): It is the proportion of events that are being tested for which yield negative test outcome with the test. It is also referred as Miss Rate.

FNR = FN/ (FN+TP)

In situations where we have another round of tests to confirm a disease effect this measure would be a good choice.

If there is a need to identify what population of people have been affected by a particular virus this measure would be a bad choice.

1. F-measure (F score or F1 score): It is the harmonic mean of precision and recall.

F1 = 2TP/ (2TP+FP+FN)

As it is the mean of precision and recall it is widely use in natural language processing. Example: Word segmentation.

1. Mathew’s correlation coefficient (MCC): it is the measure of the quality of Binary classification.

MCC = [(TP \* TN) – (FP \* FN)] / Sqrt [(TP+FP) (TP+FN) (TN+FP) (TN+FN)]

This is good for deciding quality of algorithm where outcome is like true/false or yes/no.

eg. whether person has disease or not.

This is not efficient where classes are more than two. eg. Prediction of weather where classes can be cloudy, rainy, sunny or snow.

1. False Discovery Rate (FDR): It is the ratio of number of false positives to the total number of positive outcomes.

FDR = FP/ (TP+FP) = 1-Precision (PPV)

In case of testing whether a player throws a ball into the basket is a good choice for False Discovery rate.

In case of checking whether a mail is spam or not is a bad choice for False Discovery rate.

1. False Omission Rate (FOR): It is the ratio of number of false negatives to the total number of negative outcomes.

FOR = FN/ (TN+FN) = 1-NPV

In case of a disease, doctor is predicting he doesn’t have disease is a bad example for this measure.

Where doctor is predicting that a patient has disease is a good example for this measure.

1. **Identify and correct at least one mathematical typo in chapter 2.**

Solution.

In chapter 2, Accuracy formula is defined as follows:

Accuracy = *TP* + *FP / (TP* + *FP* + *TN* + *FN)*

The above expression for Accuracy is wrong because instead of having numerator as (TP+TN) it is having (TP+FP)

Corrected expression for Accuracy is = *TP* + *TN / (TP* + *FP* + *TN* + *FN)*