CSCE 689: Medical Sensing and Analytics, Fall 2017
HW 1: Metrics of ML Models

Coding: This assignment should be completed in either R or Python 2.7, your choice.

There are two sets of data for use in this homework, the probabilities generated by a machine learning method (xgboost, to be precise), and the ground truth labels for each of the cases. The two vectors (stored in one RData file, for R, or two CSV files, for Python) are in the same order. In other words, the first element in the probability vector has a ground truth label as the first element in that vector.

Deliverables:

1) Text document with answers to the questions
2) Code file with all of your source code

We want to further evaluate the difference between machine learning toolkits that generate predictions at 50% and a variable threshold method.

The questions we want to answer are:

1) What is the AUROC for this dataset (you can use built-in packages to generate this answer, but show what package and how you used it)?

2) At a 50% threshold, what are the True Positive, True Negative, False Positive, False Negative, Precision, Recall, F-score, Sensitivity, and Specificity? (you should be able to write code that generates these from the vectors, do not use pre-built metric packages)

3) Write a function that finds the optimal threshold for decision making (above which is a 1, below which is a 0). This function should have the following properties:

   Function Name: optimalROCPoint
   Input: ground truth vector, predicted probability vector, number of points to consider (default of 1000), Condition to Optimize

   Output:
   1) The TP, FP, TN, FN, Precision, Recall, Sensitivity, Specificity, TPR, FPR, and F-score for each threshold tested (for example, of the 1000)
   2) The Optimal threshold (if multiple exists, you can choose to output all of them, randomly choose one, or always output the first)

   Some things that will help simplify this function:
   Conditions to Optimize (maximize) will be limited to the list of the following:
   1) F-score
   2) Precision

3) Recall
4) Sensitivity
5) Specificity
6) TPR given a maximum FPR threshold (this threshold must be provided)
7) Minimum FPR given a minimum TPR threshold (this threshold must be provided)

3) Self-check your Function by Plotting the TPR/FPR points and comparing this plot to an ROC plot