Calculating ROC curves and AUC scores

Preliminaries:

1. Download the file 'classprobabilities.csv' from `http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/classprobabilities.csv`. The first column is an index that you do not really need. The second column gives the true class label (either 0 or 1). The third and the fourth column give the probabilistic scores for two different classifiers. The higher this probability, the more certain the classifier is that the example belongs to class 1 (instead of class 0).

2. Download the paper 'ROC Graphs: Notes and Practical Considerations for Researchers' by Tom Fawcett from `http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/ROC101.pdf` and read it. It contains all you need to know to solve this exercise, except that the method explained for computing the area under the curve is quite complicated. A simpler formula is:
$$\text{AUC} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{p_i > p_j} \; .$$

   Here i runs over all m data points with true label 1, and j runs over all n data points with true label 0; p_i and p_j denote the probability score assigned by the classifier to data point i and j, respectively. $\mathbf{1}$ is the indicator function: it outputs 1 iff the condition (here $p_i > p_j$) is satisfied.

Technically you can do this exercise by hand, but you are strongly advised to write a computer program instead (in any language you like).

Exercise:

1. Calculate the ROC curves for the classifiers and plot them. Comment on the obtained results. Draw the ROC curve that would correspond to the performance of a classifier that assigns the class labels to the test data points randomly. Do both classifiers perform better than this baseline?

2. Measure the AUC scores (area under the curve) of both classifiers using the formula given above. Do the computed AUC scores indicate that the classifiers are performing better than this baseline?

3. Choose a threshold of 0.5, translate the probability scores to predicted class labels, and compute the accuracy for each of the classifiers. Use a statistical significance test to tell whether the performance of the two classifiers is significantly different. Here the paper by Salzberg, from `http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/salzberg97comparing.pdf`, comes in handy.

Index	True label	Predictive probability (1)	Predictive probability (2)
⋮	⋮	⋮	⋮
60	1	0.43013	0.63192
61	1	0.52135	0.52524
62	1	0.85277	0.55801
63	1	0.68814	0.71352
64	0	0.034719	0.18274
65	0	0.30511	0.75859
⋮	⋮	⋮	⋮

Table 1: Predicted scores from classifier 1 and classifier 2 and the actual class labels. For the complete dataset, see `http://www.cs.ru.nl/~tomh/onderwijs/dm/dm_files/classprobabilities.csv`.