Exercise (ROC curve and AUC)

Let us consider a binary classification problem where all $x \in \mathscr{X}$ of some state space $\mathscr{X}$ has some label $l(x) \in \{0,1\}$. We would like to find a way to compare different classifiers for this problem and to select the one that best suits our expectations. In this exercise, we explain how the *Receiver Operating Characteristic* (ROC) curves of the different classifiers can be used to make a choice.

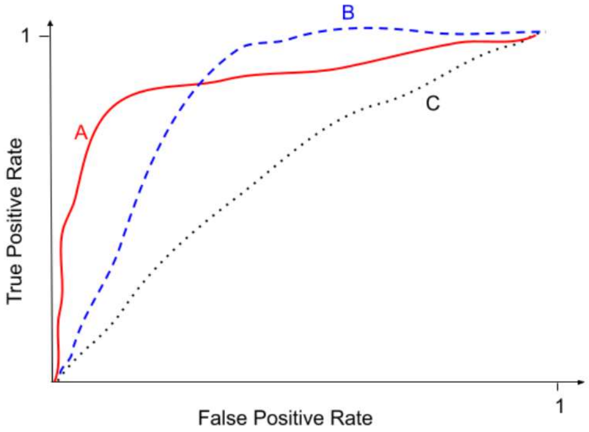
A *Receiver Operating Characteristic* (ROC) curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP + FN)). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN + FP)). For example, in medical testing, the true positive rate is the rate in which people are correctly identified to test positive for the disease in question.

A discrete classifier that returns only the predicted class gives a single point on the ROC space. But for probabilistic classifiers, which give a probability or score that reflects the degree to which an instance belongs to one class rather than another, we can create a curve by varying the threshold for the score. Concretely, if our classifier is obtained from logistic regression as explained at the end of Section 5.2.1, the ROC is obtained by computing the TPR and the FPR for every threshold value $p_0 \in [0,1]$ (with the notations of Section 5.2.1).

1. What would be the ROC curve of a "perfect" classifier (namely a classifier that never makes mistakes)?

2. What would be the ROC curve of a random classifier (namely a classifier attributing to each sample a label $l \in \{0,1\}$ by sampling l from a Bernoulli distribution with parameter 1/2)?

3. We provide with this figure ROC curves for three different classifiers (namely A, B and C). Based on this graph, can we state which is the best classifier to use? If not, can we at least get rid of one classifier that would perform always less well?



4. Given two classifiers, it is likely that the ROC curve of the first will not always be above (or below) the one of the second. In this case, it is not obvious to say which classifier should be preferred. In such situation, a standard approach consists in selecting the classifier with the largest *Area Under the Curve* (AUC). The AUC is defined as the integral of the ROC curve over $[0,1]$.
   What is the value of the AUC for respectively the perfect and the random classifiers?

5. In the machine learning community, the use of AUC remains questioned. One problem of the AUC is that it does not take into account the cost of false positives or false negatives. One advantage presented by ROC curves is that they aid us in finding a classification threshold that suits our specific problem.

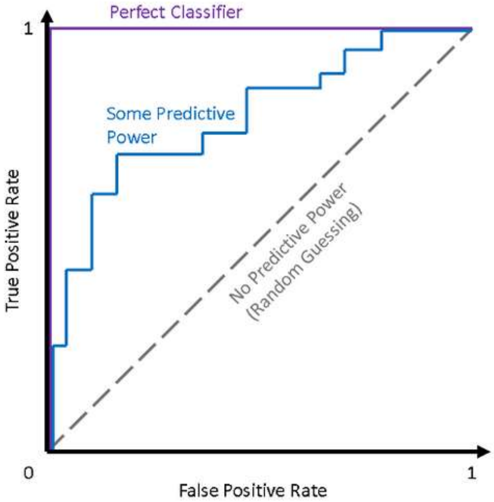
   (a) Suppose that you are evaluating an email spam classifier. You do not want someone to lose an important email to the spam filter just because your algorithm was too aggressive. You state as Positive an email that is a spam. In this situation, which classifier (among A, B and C) will you use?

   (b) Suppose now that our classifier is predicting whether someone has a terminal illness and your colleague decided to declare as Positive a sick patient. You might accept to incorrectly diagnosing the illness, just to make sure that we don't miss people who actually have the illness. In this situation, which classifier will you use?

Solutions

1. A perfect classifier would have sensitivity and specificity both equal to 1. If a cut-off value existed to produce such a test, then the sensitivity would be 1 for any non-zero values of 1 – specificity. The ROC curve would start at the origin (0,0), go vertically up the y-axis to (0,1) and then horizontally across to (1,1).

2. The random classifier is equally likely to produce a false positive or a true positive. As a result, its ROC curve is a line from $(0,0)$ to $(1,1)$. In this figure, we visualize the ROC curves for both the perfect classifier and the random classifier.



3. The ROC curve of the classifier is always below the one of the two other classifiers. This means that it is always preferable to work with the classifier A or B rather than the classifier C. Hence, we can get rid of the classifier C.

4. The AUC value of the perfect classifier is 1, while the one of the random classifier is 1/2.

5. (a) Our priority is to have a small false positive rate. As a consequence, we are encline to use classifier A rather than B.

   (b) In this case, we might accept a higher number of false positives (incorrectly diagnosing the illness), just to make sure that we don't miss any true positives (people who actually have the illness). Thus we want to reach a small False Negative Rate which is equal to one minus the True Positive Rate. Thus, we want a large True Positive Rate and we will prefer classifier B rather than A.

   To conclude this question, the take-home message is that we should pick a threshold depending on the trade off that we want to make between (i) the cost of failing to detect positives and (ii) the cost of raising false alarms.