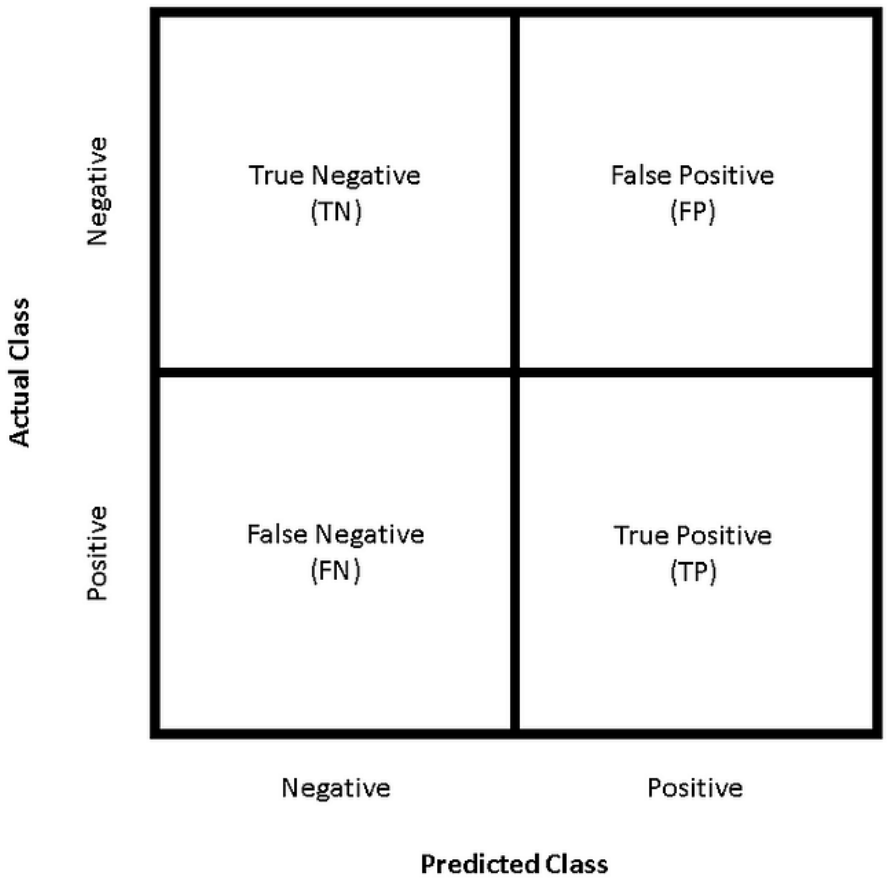
The ROC Curve

The **receiver operating characteristic (ROC) curve** is frequently used for evaluating the performance of binary classification algorithms. It provides a graphical representation of a classifier's performance, rather than a single value like most other metrics.

First, let's establish that in binary classification, there are four possible outcomes for a test prediction: **true positive**, **false positive**, **true negative**, and **false negative**.



Confusion matrix structure for binary classification problems

The ROC curve is produced by calculating and plotting the **true positive rate** against the **false positive rate** for a single classifier at a variety of **thresholds**. For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class. Normally in logistic regression, if an observation is predicted to be positive at > 0.5 probability, it is labeled as positive. However, we could really choose any threshold between 0 and 1 (0.1, 0.3, 0.6, 0.99, etc.) — and ROC curves help us visualize how these choices affect classifier performance.

The **true positive rate**, or **sensitivity**, can be represented as:

$$TPR = \text{Sensitivity} = \frac{TP}{TP + FN}$$

where **TP** is the number of **true positives** and **FN** is the number of **false negatives**. The true positive rate is a measure of the probability that an *actual* positive instance will be classified as positive.

The **false positive rate**, or 1 — **specificity**, can be written as:

$$FPR = 1 - Specificity = \frac{FP}{FP + TN}$$

where **FP** is the number of **false positives** and **TN** is the number of **true negatives**. The false positive rate is essentially a measure of how often a "false alarm" will occur — or, how often an *actual* negative instance will be classified as positive.

Figure 1 demonstrates how some theoretical classifiers would plot on an ROC curve. The gray dotted line represents a classifier that is no better than random guessing — this will plot as a diagonal line. The purple line represents a perfect classifier — one with a true positive rate of 100% and a false positive rate of 0%. Nearly all real-world

examples will fall somewhere between these two lines — not perfect, but providing more predictive power than random guessing. Typically, what we're looking for is a classifier that maintains a high true positive rate while also having a low false positive rate — this ideal classifier would "hug" the upper left corner of Figure 1, much like the purple line.

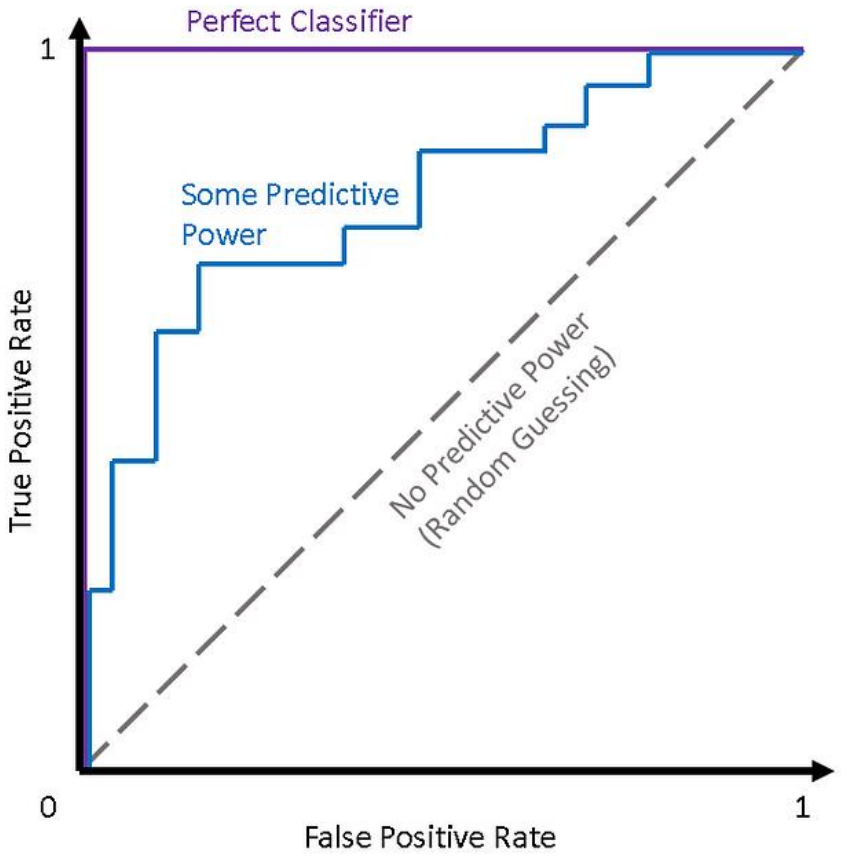


Fig. 1 — Some theoretical ROC curves

AUC

While it is useful to visualize a classifier's ROC curve, in many cases we can boil this information down to a single metric — the AUC.

AUC stands for **area under the (ROC) curve.** Generally, the higher the AUC score, the better a classifier performs for the given task.

Figure 2 shows that for a classifier with no predictive power (i.e., random guessing), AUC = 0.5, and for a perfect classifier, AUC = 1.0. Most classifiers will fall between 0.5 and 1.0, with the rare exception being a classifier performs *worse* than random guessing (AUC < 0.5).

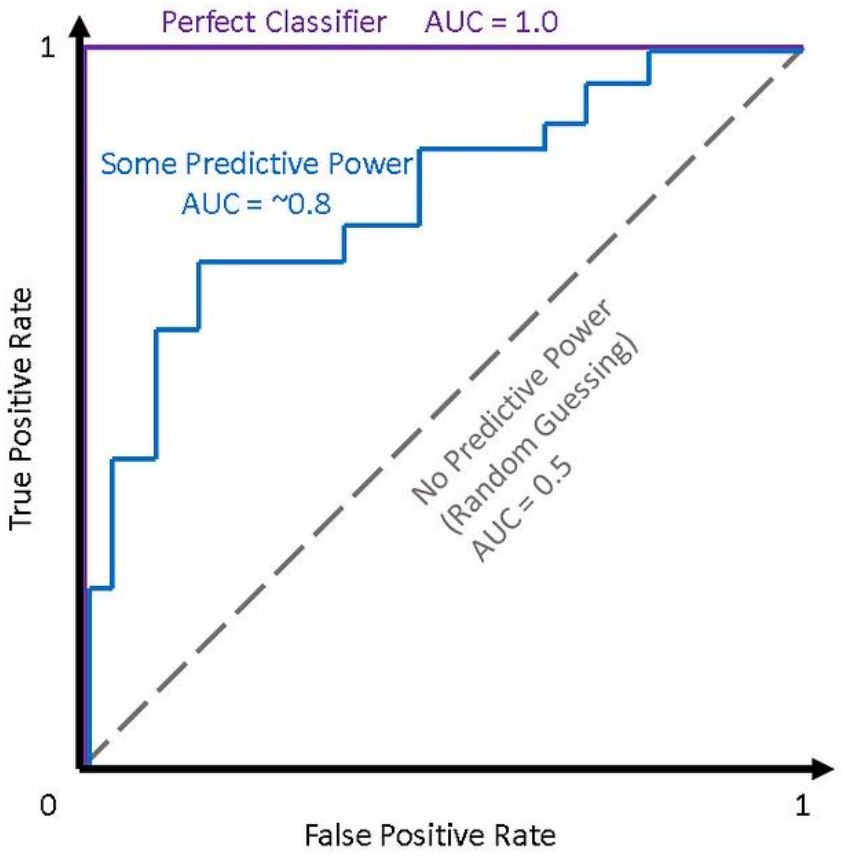


Fig. 2 — Theoretical ROC curves with AUC scores

Why use ROC Curves?

One advantage presented by ROC curves is that they aid us in finding a classification threshold that suits our specific problem.

For example, if we were evaluating an email spam classifier, we would want the false positive rate to be really, really low. We wouldn't want someone to lose an important email to the spam filter just because our algorithm was too aggressive. We would probably even allow a fair amount of actual spam emails (true positives) through the filter just to make sure that no important emails were lost.

On the other hand, if our classifier is predicting whether someone has a terminal illness, we might be ok with 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).

Additionally, ROC curves and AUC scores also allow us to compare the performance of different classifiers for the same problem.