Receiver operator characteristic - ROC

Classifying is hardly a perfect process. There will be errors depending on how interleaved (overlapped) the classes are and what do we select as the decision boundary. Putting the selection boundary in the middle of two classes is also not a straightforward choice and we need to discuss terms such as *cost* and *risk*.

Last week, we discussed classifying salmons and seabasses. We assumed that both outcomes were as desirable. However, if we were to work in a fish-packing company, this point of view would not work. The customers would probably be happy if they got salmon instead of seabass, but they would be enraged if they got a seabass instead of a salmon. Therefore, we would like to decrease the possibility for misclassifying seabasses as salmons. Classifying too many seabasses as salmons would be costly. Of course, classifying all the salmons as seabasses would be costly too as salmon is more expensive. Finding the right balance is sometimes tricky.

In decision-theoretic terminology, an expected loss is called a risk. We can define conditional risk as

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

in which λ is a loss function and α describes the possible actions. The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i when the state of nature is ω_j. That is, how much we loose by making mistakes. So, we should find a way to minimize the risk. (Book has the mathematical proofs for the subject.)

So, how do we tell how good our classifier is? How can we compare which classifier is better?

One place to start is to build a *confusion matrix*. The name stems from the fact that it makes it easy to see if the system is confusing two classes.

Total population	Predicted condition positive	Predicted condition negative
Condition positive	True positives	False negatives
Condition negative	False positives	True negatives

All the values have multiple names, such as
True positive (TP) (hit)
True negative (TN) (correct rejection)
False positive (FP) (false alarm)
False negative (FN) (miss)

From the confusion matrix, we can determine other useful values:

Sensitivity (recall, hit rate, true positive rate (TPR)) measures the proportion of positives that are correctly identified as such

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Specificity (true negative rate, TNR) measures the proportion of negatives that are correctly identified as such.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

Precision, positive predictive value

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

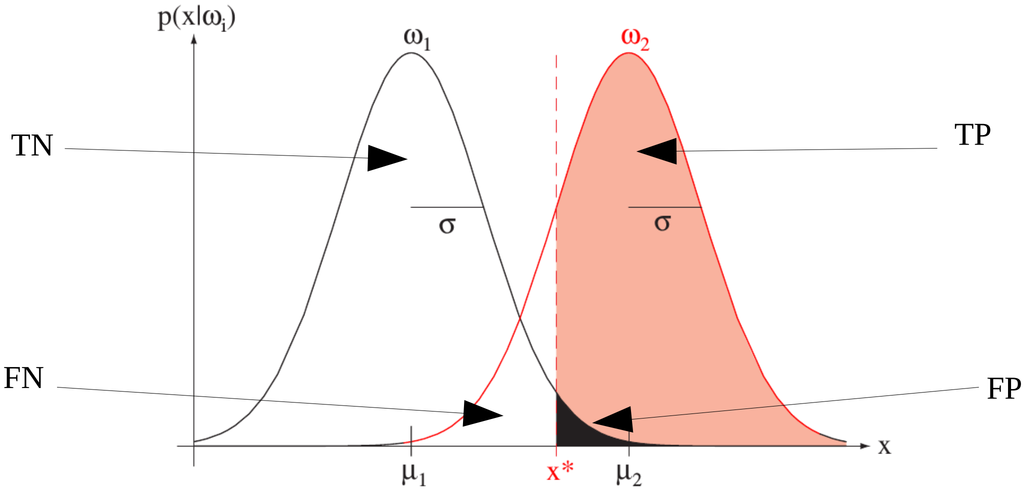
Negative predictive value

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

Accuracy

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N}$$

In the case of one feature, we can illustrate the process of selecting the decision threshold with the following image.



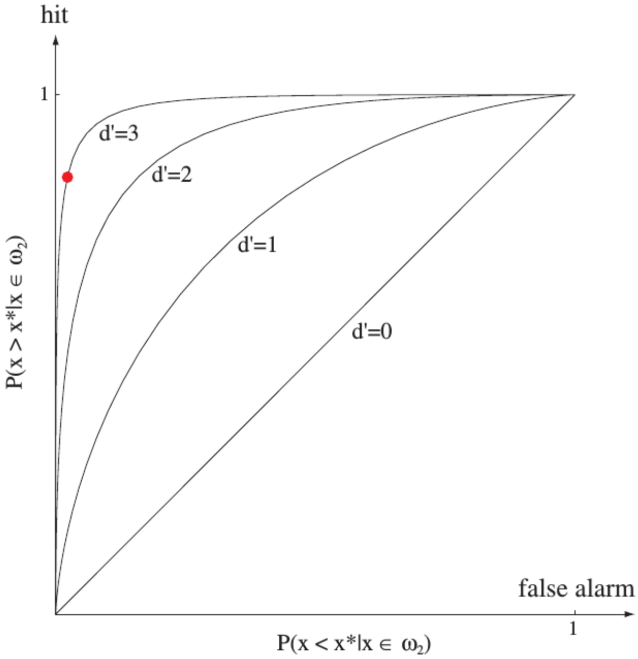
The *discriminability* of the two curves/signals is presented as

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

Notice, that at this point, we are more interested in the areas beneath the curves than the curves them selves. We wish to determine the optimal point for the decision threshold x*. How we determine what is optimal, depends on the application.

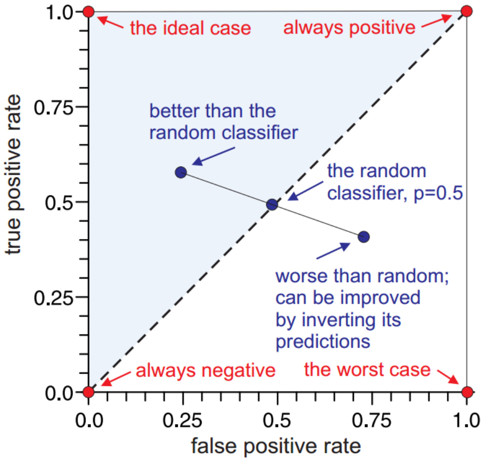
If we assume that x* is fixed but not known, we can draw a curve with sensitivity (hit rate) on one axis and 1-specificity (false alarm rate) on the other. This curve is called ROC (Receiver operating characteristic) curve.

The ROC curve can be generated by plotting the cumulative distribution function of the detection probability in the y-axis versus the cumulative distribution function of the false alarm probability in x-axis (Wikipedia).



The ROC curve can be used to choose the best operating point. The best operating point might be chosen so that the classifier gives the best trade off between the costs of failing to detect positives against the costs of raising false alarms. These costs need not be equal, however this is a common assumption.

If this was a cancer test, we would like to have higher sensitivity (fewer missed cancers). However, it will result on lower specificity and more false positives or “cancer scares”. Of course, it is much worse to miss a cancer than to endure an unnecessary biopsy, so we tend to choose a higher sensitivity cutoff in real world practice.


1.0
0.75
0.5
0.25
0.0
true positive rate
the ideal case
always positive
better than the random classifier
the random classifier, p=0.5
worse than random; can be improved by inverting its predictions
always negative
the worst case
0.0
0.25
0.5
0.75
1.0
false positive rate