Precision recall curves

Peter Corke

December 2016

1 Binary classifiers

Binary classifiers are widely used in fields as diverse as document retrieval and robot navigation. We will first consider a familiar case of document retrieval using a tool like Google where we perform a search and expect to see only relevant documents. The classifier, Google in this case, will:

- return a relevant document (true positive)
- return an irrelevant document (false positive)
- choose not to return a relevant document (false negative)
- chooses not to return an irrelevant document (true negative)

If we consider that the classification of the document is either positive (relevant) or negative (not relevant) then we can express these four outcomes in a 2×2 contingency table or confusion matrix as shown in Figure 1. We introduce the common shorthand notation TP, FP, FN and TN. The two error situations FP and FN are often referred to as type I and type II errors respectively. False positives or type I errors are also referred to as false alarms. The number of relevant documents is $TP + FN$ and the number of irrelevant documents is $TN + FP$.

Consider now a robot localization problem. The robot is at a particular point X and its localizer can do one of four things:

- correctly report the robot is at X (true positive)
- incorrectly report the robot is at X (false positive)
- incorrectly report the robot is not at X (false negative)
- correctly report the robot is not at X (true negative)

For robot navigation we need to determine what it means to be "at location X" given the inevitable errors in sensor data that lead to uncertainty in location. Typically there is some threshold test that depends on the scale of the environment. For FABMAP[1] and SeqSLAM[2] (and its derivatives) this is 40 m.

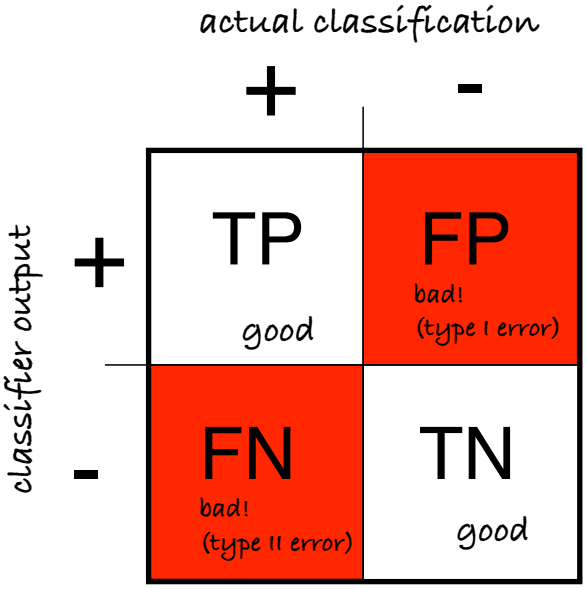

Figure 1: Confusion matrix for a binary classifier which maps the true classification against the output of the classifier.

1.1 Accuracy

Accuracy is the number of correct classifications as a fraction of all classifications. It is the sum of the diagonals of the confusion matrix divided by the total

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

However for the case of imbalanced classes, that is, where one of the classes is very unlikely this measure can be very misleading. Consider the case of classifying airline passengers as a terrorist (T) or not-a-terrorist (NaT) [3]. For the period 2000–17 there were 800M passengers per year and only 19 confirmed terrorists. A classifier that declared everybody as NaT would have

TP	T classified as T	0
TN	NaT classified as NaT	$18 \times 800 \times 10^6$
FP	NaT classified as T	0
FN	T classified as NaT	19

Accuracy in this case is

$$\frac{18 \times 800 \times 10^6}{18 \times 800 \times 10^6 + 19} \approx 99.9999999\%$$

which is very good, but misses the point of being useful since it would not detect a single terrorist.

2 Precision and recall

For a particular experiment (over a number of documents or a number of points along the robot's path) where we know the true classification, we can compute the number of instances of each element in the table shown in Figure 1. From these four numbers we can compute a number of performance measures:

- Precision is the fraction of retrieved documents that are relevant, or the fraction of localizations that are correct

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

- Recall is the fraction of relevant documents that are retrieved, or the fraction of correct localizations that are given

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

A classifier with high precision gives very few incorrect answers. For information retrieval this means very few irrelevant documents are returned, while for localization this means very few incorrect locations are returned. We could imagine a classifier which achieves this by being very conservative, that is, it would avoid returning documents that it is not completely certain about and therefore fails to return many relevant documents. For localization this is equivalent to reporting a location infrequently, that is, only when it is very certain. Such a classifier would make a large number of false negative (type II) errors, and we see from the equation for precision that these are not accounted for. Precision alone is not enough to describe the performance of a classifier.

A classifier with high recall misses very few correct answers. For information retrieval this means very few relevant documents are missed, while for localization this means very few locations are not reported. We could imagine a classifier which achieves this by being very permissive, that is, it would return documents that it is not completely certain about and therefore also return many irrelevant documents. For localization this is equivalent to reporting a location very frequently, even if it is not very certain. Such a classifier would make a large number of false positive (type I) errors, and we see from the equation for recall that these are not accounted for. An ideal classifier would therefore have both high precision and high recall.

We mentioned above about the classifier being conservative or permissive but in fact classifiers lie along a continuous spectrum between these two extremes. A binary classifier's output is based on the input data and some decision boundary (or discrimination threshold) which might be a simple threshold test or the margin in a support vector machine (SVM). As we adjust the threshold the precision and recall of the classifier will vary.

A precision-recall (PR) curve is simply a plot of recall against precision for a given experiment as the decision threshold is varied. An example PR curve is shown in Figure

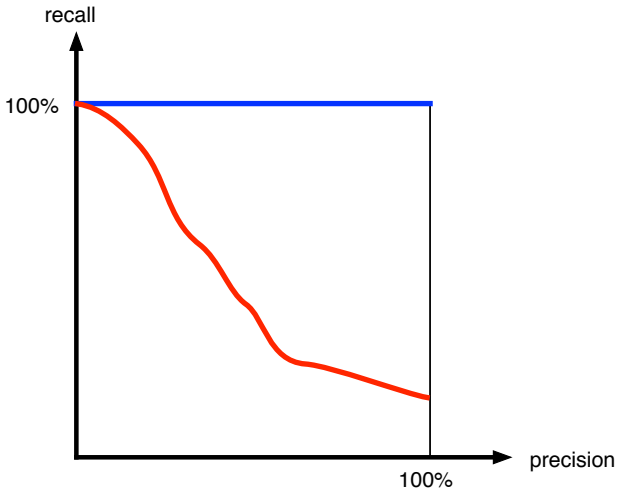

Figure 2: Precision-recall (PR) curves. In blue is the performance of an ideal classifier and in red is the performance of an imperfect classifier.

3. The ideal classifier is shown in blue – at 100% precision (perfect classification) it has 100% recall (every positive is returned). A more realistic classifier is shown in red. At 0% precision it returns 100% of correct classifications but also a huge number of misclassifications. At 100% precision it returns only a small fraction of correct classifications. In practice the PR curve would be computed for a certain number of threshold values and the results interpolated.

For the aircraft passenger screening example above we can compute precision and recall

$$P = \frac{0}{0+0} = 0, \quad R = \frac{0}{0+19} = 0$$

which are both zero (technically P is undefined).

2.1 Scalar performance measures

We have already discussed how precision and recall alone are insufficient to describe the performance of a classifier. A useful scalar performance measure is the F_1 score which is the harmonic mean of precision and recall

$$F_1 = 2\frac{PR}{P+R} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

and a high number is best.

For a PR curve a useful scalar measure is the AUC (area under the precision-recall curve) score. For an ideal classifer (the blue line in Figure 3) this would be equal to one, or for a real classifier it would be less than one.

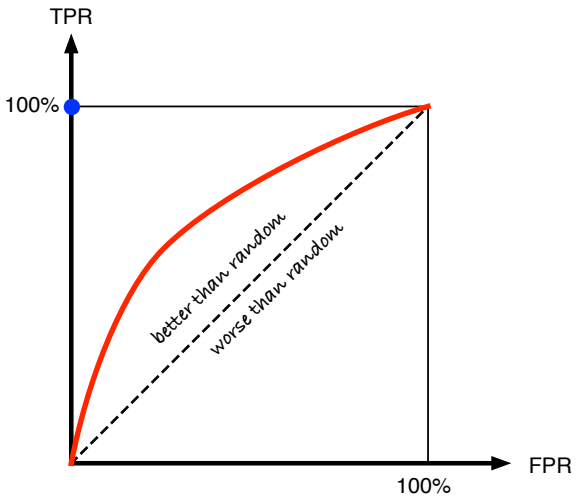

Figure 3: Receiver operating characteristic (ROC) curves. In blue is the performance of an ideal classifier – a point. In red is the performance of an imperfect classifier. The diagonal line represents the performance of a classifier that makes a random decision, effectively tossing a coin.

In the deep learning world "average precision" scalar measures are used, for example mean average precision or mAP. There are quite a few variants but they essentially take an average of the precision values for a range of recall values.

3 Receiver operating characteristic (ROC)

A closely related concept to the PR curve is the ROC curve which is based on other measures computed from the confusion matrix:

- Sensitivity or true positive rate

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

- Specificity or true negative rate

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

- Fallout or false positive rate

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

The first two, sensitivity and specificity, are commonly used to describe the performance of medical tests. A plot of TPR againt FPR is known as a receiver operating characteristic (ROC) curve[1] which is ideally a single point. The diagonal line is the performance of a random classifier or coin tosser.

There is some argument that PR curves are better for the case where the probability of positives and negatives in the input data, the priors, are important.

References

[1] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[2] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE International Conference on Robotics and Automation*, pp. 1643–1649, May 2012.

[3] W. Koehrsen, "Beyond accuracy: Precision and recall."

[1] Developed during WW2 for analysis of radar signals to detect aircraft.