Receiver operating characteristics (ROC) graphs in classification

Jordi Villà i Freixa

Universitat de Vic - Universitat Central de Catalunya
Study Abroad

jordi.villa@uvic.cat

course 2023-2024

Índex

What to expect?

In this session we will discuss:

- Classifier performance
- ROC space
- Generation of ROC curves
- Area under the curve (AUC)

- A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their preformance.
- Simple classification accuracy is a poor meric for measuring performance,
- In addition, ROC curves have properties specially useful for skewed class distribution and unequal classification error costs.

Classifier performance

Let us start by assuming just two classes for the instances I, positive and negative: $\{\mathbf{p}, \mathbf{n}\}$. A *classification model* or *classifier* is a mapping from instances to predicted classes $\{\mathbf{Y}, \mathbf{N}\}$.

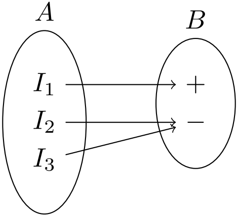


Figure 1: A classifier is a mapping between the group of instances and the group of categories or labels

Some models produce a continuous output (estimation of and instance's class membership probability) to which different thresholds may be applied to predict class membership.

Confusion matrix (or contingency table)

True class

Hypothesized class	**p**	**n**
Y	True Positives	False Positives
N	False Negatives	True Negative
total	P	N

$$\mathrm{FPR} = \frac{FP}{N} = 1 - \mathrm{TPR}$$

$$\mathrm{TPR} = \frac{TP}{P} = 1 - \mathrm{FPR}$$

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

$$\text{recall (or sensitivity)} = \frac{\mathrm{tp}}{\mathrm{tp+fn}}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\mathrm{F}_{\text{measure}} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

sensitivity = recall = hit rate = TPR // specificity = selectivity = TNR

ROC space

An ROC grph depicts relative tradeoffs between befeits (TP) and costs (FP).

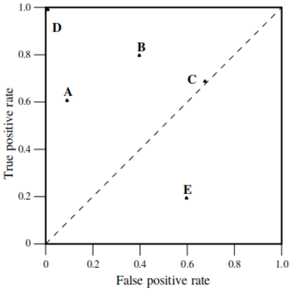


Figure 2: Several examples of a discrete classifier[1].

- $(0,0)$ strategy of never issuing a positive classification
- $(1,1)$ unconditional issuing positive classifications
- $(0,1)$ perfect classification
- $(\approx 0, \approx 0)$ *Conservative* classifiers (few errors, but strong evidence for positives)
- $(\approx 1, \approx 1)$ *Liberal* classifiers (more positive with weak evidence)

Some interesting regions

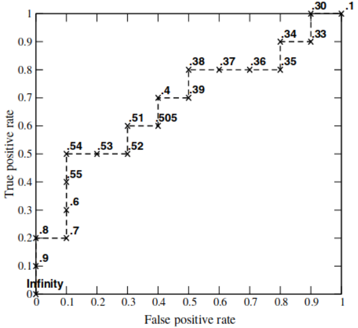
- Random performance $y = x$
- To get away from the diagonal, the classifier should exploit some information in the data.
- Any classifier that generates a point in the lower right triangle can be *negated* to produce a dot in the upper left triangle.
- The question is: is a classifier slightly better than random signifficative or is it only better than random by chance? To answer this, we move into ROC curves.

ROC curves

1. *Discrete classifiers* (decision trees or rule sets) only produce one point in the ROC space: a single confusion matrix. They can be transformed into a curve if we generate a score from the values obtained.
2. *Probabilistic classifiers* produce an instance an strict probability or an uncalibrated score (Naive Bayes or neural networks). We can set up a threshold to produce a binary (discrete) classifier $\{\mathbf{Y}, \mathbf{N}\}$.

ROC curves

Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1



The ROC curve created by thresholding a test set (adapted from [1]).

Missclassification error and accuracy

Remember than in the binary classification case ($c = 2$), and using the indicator loss function, the missclassification error can be written as:

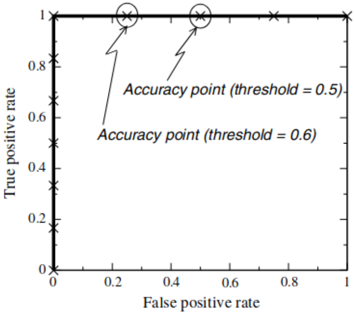
$$\text{error} = \frac{\text{FP} + \text{FN}}{P + N}$$

and the accuracy can be calculated by measuring the fraction of correctly classified objects:

$$\text{accuracy} = 1 - \text{error} = \frac{\text{TP} + \text{TN}}{P + N}$$

ROC graphs measure the ability of a classifier to produce good relative instance scores, able to discriminate between positive and negative instances.

Relative vs absolute scores



Inst	Class		Score
no.	True	Hyp	
1	p	Y	0.99999
2	p	Y	0.99999
3	p	Y	0.99993
4	p	Y	0.99986
5	p	Y	0.99964
6	p	Y	0.99955
7	n	Y	0.68139
8	n	Y	0.50961
9	n	N	0.48880
10	n	N	0.44951

Figure 3: Accuracy vs ROC: score (not properly calibrated) and classification of 10 Naive Bayes instances, and the resulting ROC curves[1].

Precision-Recall curves

The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

As deduced from Slide 6, ROC curves are insensitive to changes in class distribution: if the proportion of positive to negative instances changes in a test set, the ROC curves will not change! Accuracy, precision or F score are sensitive to class skews.

Precision-Recall curves

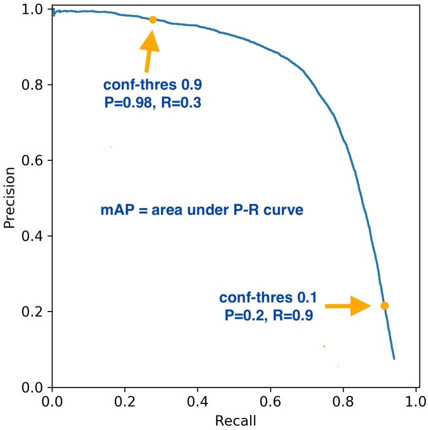


Figure 4: Example of precision-recall curve.

Class skew

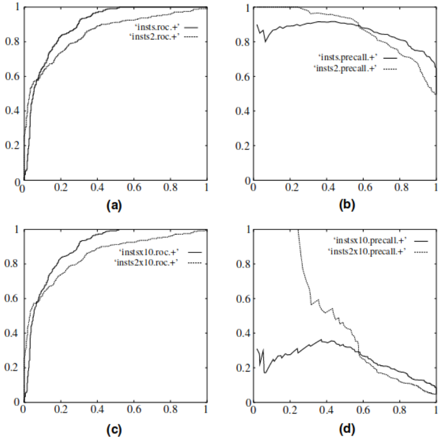


Figure 5: ROC and precision-recall curves under class skew. a-b) 1:1 rates; c-b) 1:10 rates; a-c) ROC curves; b-c) PR curves[1].

Convex hull

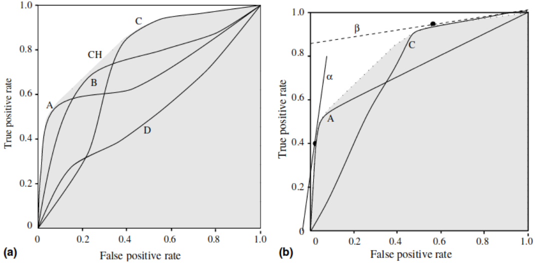
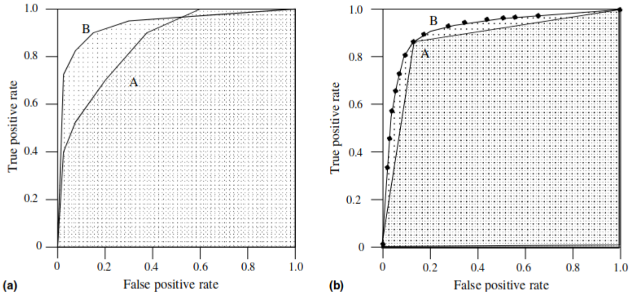


Figure 6: (a) Potentially optimal classifiers from ROC curves. Isoperformance line: $\frac{TP_2 - TP_1}{FP_2 - FP_1} = m$ for points with same expected cost. (b) Lines α and β show the optimal classifier under different sets of conditions[1].

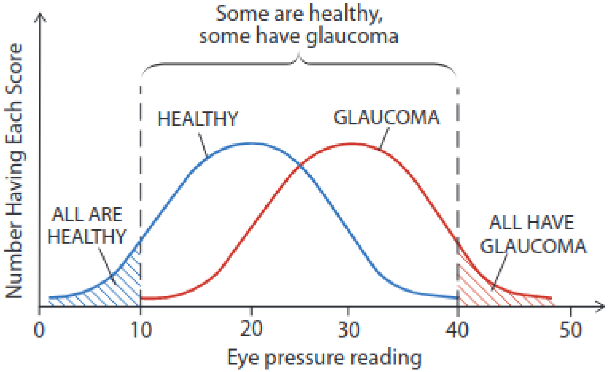
Area under the curve (AUC)

The AUC of a classifier is equivalent to the probability that the classifier will rank a randonmly chosen positive instance higher than a randomly chosen negative instance (equivalent to Wilcoxon test of ranks)[1].



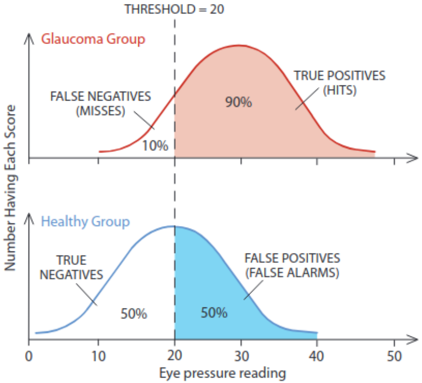
A complete example[2]

STEP 1: sample population of people whose eye pressure level and glaucoma status is known.



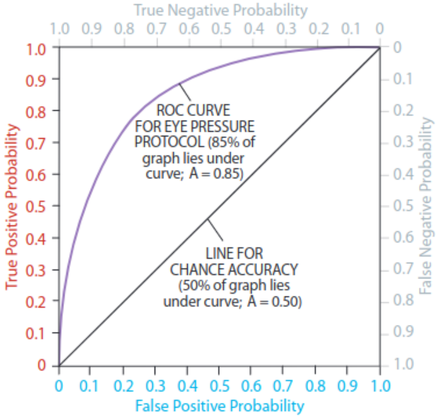
A complete example[2]

STEP 2: determine the fraction of patients in the same population who would have properly diagnosed if a given threshold was applied



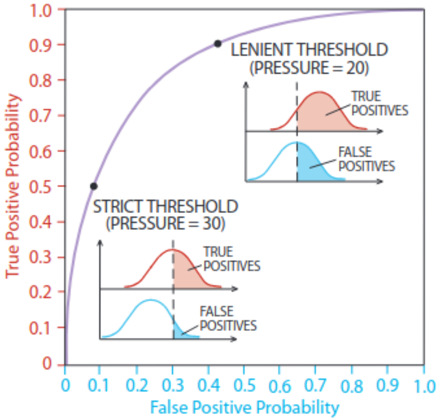
A complete example[2]

STEP 3: build a ROC curve for the different threshold values



A complete example[2]

STEP 4: select a threshold for yes/no diagnoses. Threshold chosen may often depend on subjective factors.



Practical implementation in python

Many examples of practical implementation of a ROC and precision recall curves in python are available. See, e.g., this example.

Tom Fawcett.
An introduction to ROC analysis.
Pattern Recognition Letters, 27(8):861–874, June 2006.

John A. Swets, Robyn M. Dawes, and John Monahan.
Better Decisions through Science.
Scientific American, 283(4):82–87, October 2000.