

Understanding ROC curve

Asked 10 years, 1 month ago Modified 6 years, 4 mo



81



I'm having trouble understanding the ROC

Is there any advantage / improvement in from each unique subset of the training :

has values of $\{a, a, a, a, b, b, b, b\}$, and I build model A by using a from 1st-4th values of y and 8th-9th values of y and build model B by using remained train data. Finally, generate probability.

Any thoughts / comments will be much appreciated.

Here is r code for better explanation for my question:

```
Y = factor(0,0,0,0,1,1,1,1)
X = matrix(rnorm(16,8,2))
ind = c(1,4,8,9)
ind2 = -ind

mod_A = rpart(Y[ind]~X[ind,])
mod_B = rpart(Y[-ind]~X[-ind,])
mod_full = rpart(Y~X)

pred = numeric(8)
pred_combine[ind] = predict(mod_A,type='prob')
pred_combine[-ind] = predict(mod_B,type='prob')
pred_full = predict(mod_full, type='prob')
```

So my question is, area under ROC curve of `pred_combine` vs `pred_full`.

`r` `roc`

Share Cite Improve this question

Follow

edited Jan 9, 2015 at 22:18



gung - Reinstate Monica
146k 89 402 711

asked Jul 2, 2014 at 7:18




Tay Shin
1,015 2 8 10

5 A better example would do a lot to improve the question. – mpiktas Jul 2, 2014 at 8:17

My understanding is that you want to increase AUC by choosing some specific samples? If that is your purpose, I strongly believe that this approach of biased sample selection is completely wrong, at least if your purpose is to find a good measure for classification performance. – rapaio Jul 2, 2014 at 8:19

1 Answer

Sorted by: Highest score (default) 



I'm not sure I got the question, but since the title asks for explaining ROC curves, I'll try.

203

ROC Curves are used to see how well your classifier can separate positive and negative examples and to identify the best threshold for separating them.



To be able to use the ROC curve, your classifier has to be *ranking* - that is, it should be able to rank examples such that the ones with higher rank are more likely to be positive. For example, [Logistic Regression](#) outputs probabilities, which is a score you can use for ranking.



Drawing ROC curve

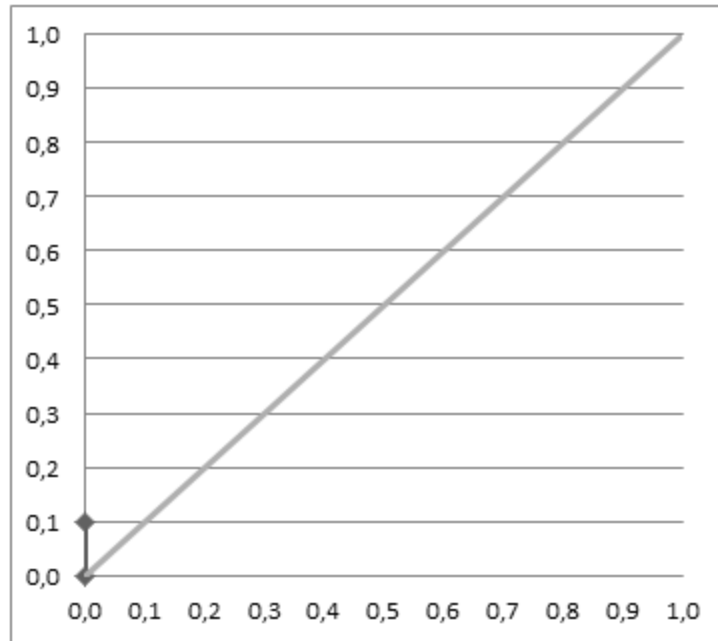
Given a data set and a ranking classifier:

- order the test examples by the score from the highest to the lowest
- start in $(0, 0)$
- for each example x in the sorted order
 - if x is positive, move $1/\text{pos}$ up
 - if x is negative, move $1/\text{neg}$ right

where `pos` and `neg` are the fractions of positive and negative examples respectively.

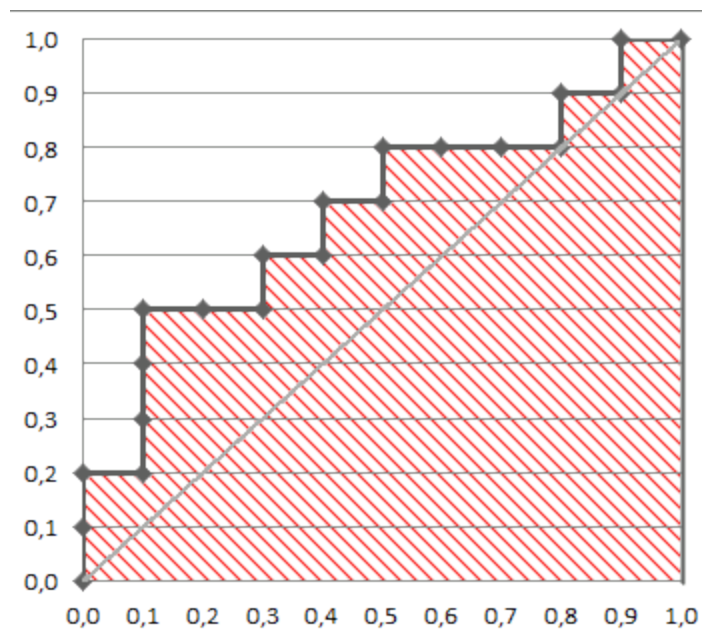
This nice gif-animated picture should illustrate this process clearer

#	C	Score
1	P	0,9
2	P	0,8
3	N	0,7
4	P	0,6
5	P	0,55
6	P	0,54
7	N	0,53
8	N	0,52
9	P	0,51
10	N	0,505
11	P	0,4
12	N	0,39
13	P	0,38
14	N	0,37
15	N	0,36
16	N	0,35
17	P	0,34
18	N	0,33
19	P	0,3
20	N	0,1



On this graph, the y -axis is true positive rate, and the x -axis is false positive rate. Note the diagonal line - this is the baseline, that can be obtained with a random classifier. The further our ROC curve is above the line, the better.

Area Under ROC



The area under the ROC Curve (shaded) naturally shows how far the curve from the base line. For the baseline it's 0.5, and for the perfect classifier it's 1.

You can read more about AUC ROC in this question: [What does AUC stand for and what is it?](#)

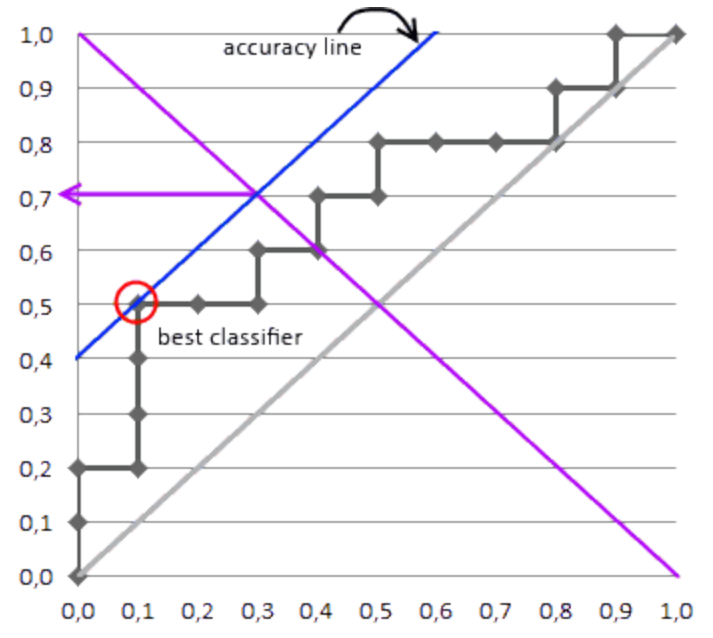
Selecting the Best Threshold

I'll outline briefly the process of selecting the best threshold, and more details can be found in the reference.

To select the best threshold you see each point of your ROC curve as a separate classifier. This mini-classifier uses the score the point got as a boundary between + and - (i.e. it classifies as + all points above the current one)

Depending on the pos/neg fraction in our data set - parallel to the baseline in case of 50%/50% - you build ISO Accuracy Lines and take the one with the best accuracy.

Here's a picture that illustrates that and for details I again invite you to the reference



Reference

- http://mlwiki.org/index.php/ROC_Analysis

Share Cite Improve this answer

Follow

edited Mar 15, 2018 at 23:14



Michael R. Chernick

43k 28 85 159

answered Jul 2, 2014 at 19:20



Alexey Grigorev

8,967 3 31 41

Just curious, your step size would have to depend on the number of positive/negative labels produced by your classifier correct? I.e. In the gif, the step size upwards is .1, if you had an extra positive label (in place of a negative label), then the "curve" would end up at 1.1 on the vertical axis (or maybe I am missing something?). So, in that case your step size needs to be smaller? – [killajoule](#) Mar 8, 2015 at 15:50

2 @gung understood. Alexey : instead of positive and negative examples, I think it should be: true positives and false positives. You may be able to see my edition of the answer, which was reverted by gung. thanks – [Escachator](#) Jun 15, 2016 at 16:05

2 I guess this ambiguity would be solved by labeling the axes (which would anyways be a good idea - but particularly for an answer that is to explain the graph) – [cbeleites](#) Jun 15, 2016 at 21:15

- 3 @AlexeyGrigorev, love the reply you give and vote up. I am not sure if there are two ROC definitions. I am referring to the ROC definition here (en.wikipedia.org/wiki/Receiver_operating_characteristic), the x-axis should be false positive rate, which is (# of predictions to be positive, but should be negative) / (# of total negative), I think in the reference, I think the x-axis is not drawing false positive rate, which does not consider the (# of total negative)? – Lin Ma Aug 28, 2016 at 22:50
-
- 3 Is there any chance of the axis being labelled in these plots? – baxx Apr 1, 2019 at 22:06
-