Pragmatic Programming Techniques

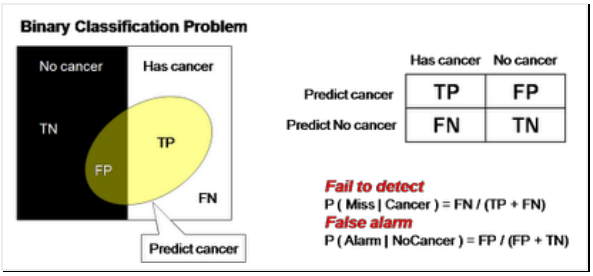
Saturday, March 19, 2011

Compare Machine Learning models with ROC Curve

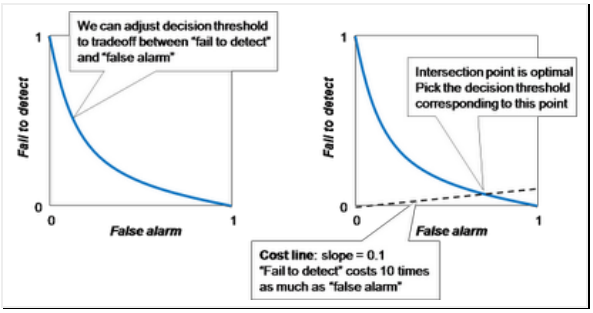
ROC Curve is a common method to compare performance between different models. It can also be used to pick trade-off decisions between "false positives" and "false negatives". ROC curve is defined as a plot of "false positive rate" against "false negative rate". However, I don't find the ROC concept is intuitive and has been struggled for a while to grasp the concept.

Here is my attempt to explain ROC curve from a different angle. We use a binary classification example to illustrate the idea. (ie: predicting whether a patient has cancer or not)

First of all, all predictive model is not 100% correct. The desirable state is that a person who actually has cancer got a positive test result, and a person who actually has no cancer got a negative test result. Since the test is imperfect, it is possible that a person who actually has cancer was tested negative (ie: Fail to detect) or a person who actually has no cancer was tested positive (ie: False alarm).



In reality, there is always a tradeoff between the false negative rate and the false positive rate. People can tune the decision threshold to adjust them (e.g. In "random forest", we can set the threshold of predicting positive when more than 30% decision trees predicting positive). Usually, the threshold is set based on the consequence or cost of mis-classification. (e.g. in this example, fail to detect has a much higher cost than a false alarm)



This can also be used to compare model performance. A good model is one that has both low false positive rate and low false negative rate, which is indicated in the size of the gray area below (the smaller the better).

"Random guess" is the worst prediction model and is used as a baseline for comparison. The decision threshold of a random guess is a number between 0 to 1 in order to determine between positive and negative prediction.

About Me

Ricky Ho

I am a software architect and consultant passionate in Distributed and parallel computing, Machine learning and Data mining, SaaS and Cloud computing.

View my complete profile

Popular Posts

MongoDb Architecture
NOSQL has become a very heated topic for large web-scale deployment where scalability and semi-structured data driven the DB requirement tow...

Designing algorithms for Map Reduce
Since the emerging of Hadoop implementation, I have been trying to morph existing algorithms from various areas into the map/reduce model. ...

NOSQL Patterns
Over the last couple years, we see an emerging data storage mechanism for storing large scale of data. These storage solution differs quite...

Couchbase Architecture
After receiving a lot of good feedback and comment on my last blog on MongoDb , I was encouraged to do another deep dive on another popular ...

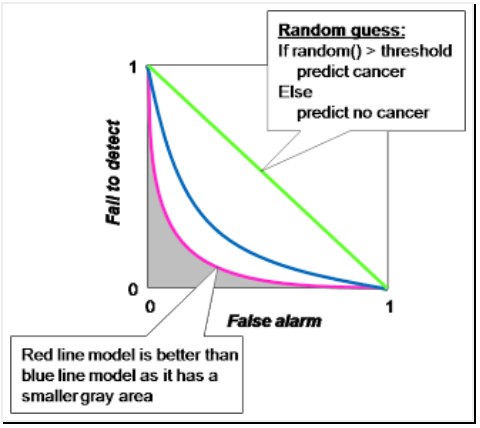
Predictive Analytics: Overview and Data visualization
I plan to start a series of blog post on predictive analytics as there is an increasing demand on applying machine learning technique to ana...

Predictive Analytics: Generalized Linear Regression
In the previous 2 posts, we have covered how to visualize input data to explore strong signals as well as how to prepare input data to a fo...

BigTable Model with Cassandra and HBase
Recently in a number of "scalability discussion meeting", I've seen the following pattern coming up repeatedly ... To make you...

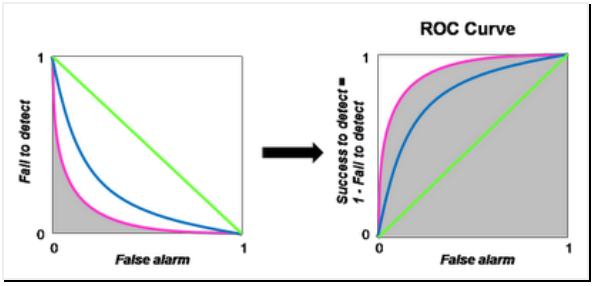
Blog Archive

► 2013 (8)
► 2012 (18)
▼ 2011 (6)
► September (1)



ROC Curve is basically what I have described above with one transformation, which is transforming the y-axis from "fail to detect" to 1 - "fail to detect", which now become "success to detect". Honestly I don't understand why this representation is better though.

Now, the ROC curve will look as follows ...



Posted by Ricky Ho at 6:47 PM

+1 Recommend this on Google

Labels: data mining, machine learning, ROC

3 comments:

Fivor said...

Hello, Ricky Ho, you have a very interesting blog for me, in particular the articles on architectural solutions for heavy web projects. Could you recommend some good books on architecture, please. For articles on datamining, machine learning, individual thank you - now, very few bloggers who write interesting articles on this topic.

March 23, 2011 at 12:47 AM

lupino3 said...

The curve that you describe before the inversion of the y-axis is called Detection Error Tradeoff (DET): http://en.wikipedia.org/wiki/Detection_Error_Tradeoff

March 24, 2011 at 9:02 AM

Unknown said...

how can I plot a ROC of a PLS-DA model? What softwares should I use?

August 27, 2012 at 7:49 PM

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

► August (1)
► July (1)
► April (1)
▼ March (2)
Compare Machine Learning models with ROC Curve
Predictive Analytics Conference 2011
► 2010 (18)
► 2009 (31)
► 2008 (22)
► 2007 (11)

Search This Blog

Search

Labels

machine learning data mining map reduce Architecture Design Cloud computing algorithm NOSQL Hadoop scalability Distributed system parallel processing big data predictive analytics Design patterns SOA performance REST ensemble method recommendation engine

Pages

- Home