Evaluation of learned models

Adapted from material by

Kurt Driessens, Evgueni Smirnov, and Hendrik Blockeel

Evaluation of Regression Models

Common measures are MSE and RMSE

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2 \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

divide by n to give measure of estimate per example. as loss function didnt matter

Another measure is the coefficient of determination: R^2

r^2 become 0 at perfect prediction as SSres = 0
r^2 becomes 1 .. baseline worst case of predicting average - the least it should do
r^2 become negative when theres a bug

ground truth - intercept giving r^2 = 0

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

sum of differences between similar variances, ground truth and average

$$SS_{\mathrm{tot}} = \sum (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

sum square residule
sum square total

sum of difference between ground truth and predicted value

$$SS_{\mathrm{res}} = \sum_{i} (y_i - h(x_i))^2$$

Confusion Matrix

In a binary classification problem, the elements of the *confusion matrix are*:

		Predicted class		
		Pos	Neg	
Actual class	+	*TP*	*FN*	*P*
	-	*FP*	*TN*	*N*

- TP: True Positive, number of positive examples which have been correctly classified as positive
- TN: True Negative, number of negative examples which have been correctly classified as negative.
- FP: False Positive, number of negative examples which have been incorrectly classified as positive.
- FN: False Negative, number of positive examples which have been incorrectly classified as negative.

Metrics for Classifier's Evaluation

P+N = total number

Accuracy = $(TP+TN)/(P+N)$

Error = $(FP+FN)/(P+N)$

Precision = $TP/(TP+FP)$

TP rate (Recall) = TP/P

FP Rate = FP/N

Precision and Recall are good metric for *information retrieval* systems.

Where:

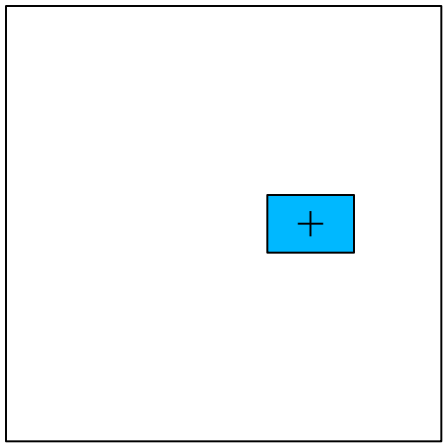
$P=TP+FN$

$N=TN+FP$

		Predicted class		
		Pos	Neg	
Actual class	+	*TP*	*FN*	*P*
	-	*FP*	*TN*	*N*

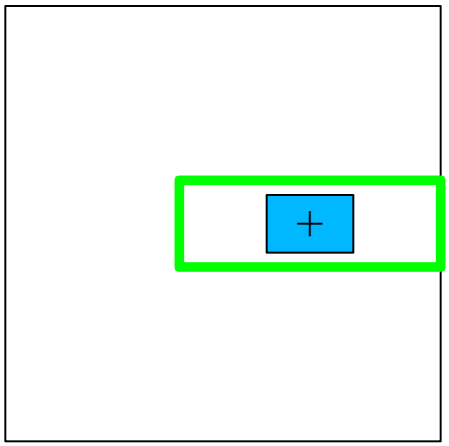
Assume the correct classification for the entire input space is “negative” except for the blue region that must be “positive”. Which classifier is better?

h(x):
    return “negative”
(i.e. always predicts as negative)



96% correct

h(x):
    if x is inside green area:
        return positive
    else:
        return negative



92% correct

If the positive class is more important or revealing patterns in data matters then classifier 2 might be actually better.

Accuracy may not be appropriate

Accuracy/Error can be misleading:

- 99% accuracy when the two classes are equally likely is good.
- 99% accuracy when the majority class is 99 times more likely, may not be so good.
- Accuracy assumes equal misclassification cost.
- Sometimes, a false negative prediction may be more costly than a false positive prediction.

Rates as estimated probabilities

- *TPR*: (true positive rate) is the estimated probability that an actual *positive* instance is classified correctly. *TPR = TP/(TP+FN) = TP/P*
- *TNR*: (true negative rate) is the estimated probability that an actual *negative* instance is classified correctly. *TNR = TN/(TN+FP) = TN/N*
- *FPR = 1-TNR*: (false positive rate) is the estimated probability that an actual *negative* instance is classified as positive *FPR = FP/(FP+TN) = FP/N*
- *FNR = 1-TPR* (false negative rate) is the estimated probability that an actual *positive* instance is classified as positive. *FNR = FN/(FN+TP) = FN/P*
- Accuracy: is the estimated probability that *some* instance is classified correctly

Misclassification Cost

- C_{FP} : cost of false positive
- C_{FN} : cost of false negative

cost of FP * prob of making a FP error + CFN * probs of making a FN error
p(posl-) = probs of predicting positive when its actually negative * p(-) probs of having a neg examples. predicting negative and the examle it is actually negative.

➔ Expected cost of a single prediction:

$$E[C] = C_{FP}\, p(pos|-)\, p(-) + C_{FN}\, p(neg|+)\, p(+)$$

how common postive are

– estimated by $C = C_{FP}\, FPR\, N/(P+N) + C_{FN}\, FNR\, P/(P+N)$

only the blue model depend on the model - a particular classifier.

Note :

the purple value tell just how common negative and positive examples there respectively
purple depend on problem, not model, as it is just ratio of negative and postive examples, not how well your model is performing.

– C is not computable from Acc or Error alone

all values can be from confusion matrix except for cost
(cost is given to you)

Expected cost is computed

Cost Sensitive Learning

Simple methods for cost sensitive learning:

- Resampling of instances according to costs

  if cost given to you can you use it to learn

- Weighting of instances according to costs

  if learing algorithm accept weights can apply it to the lost functions. just mutiple by corresponding cost (or reLATIVE COST given absolute value)

ROC diagrams

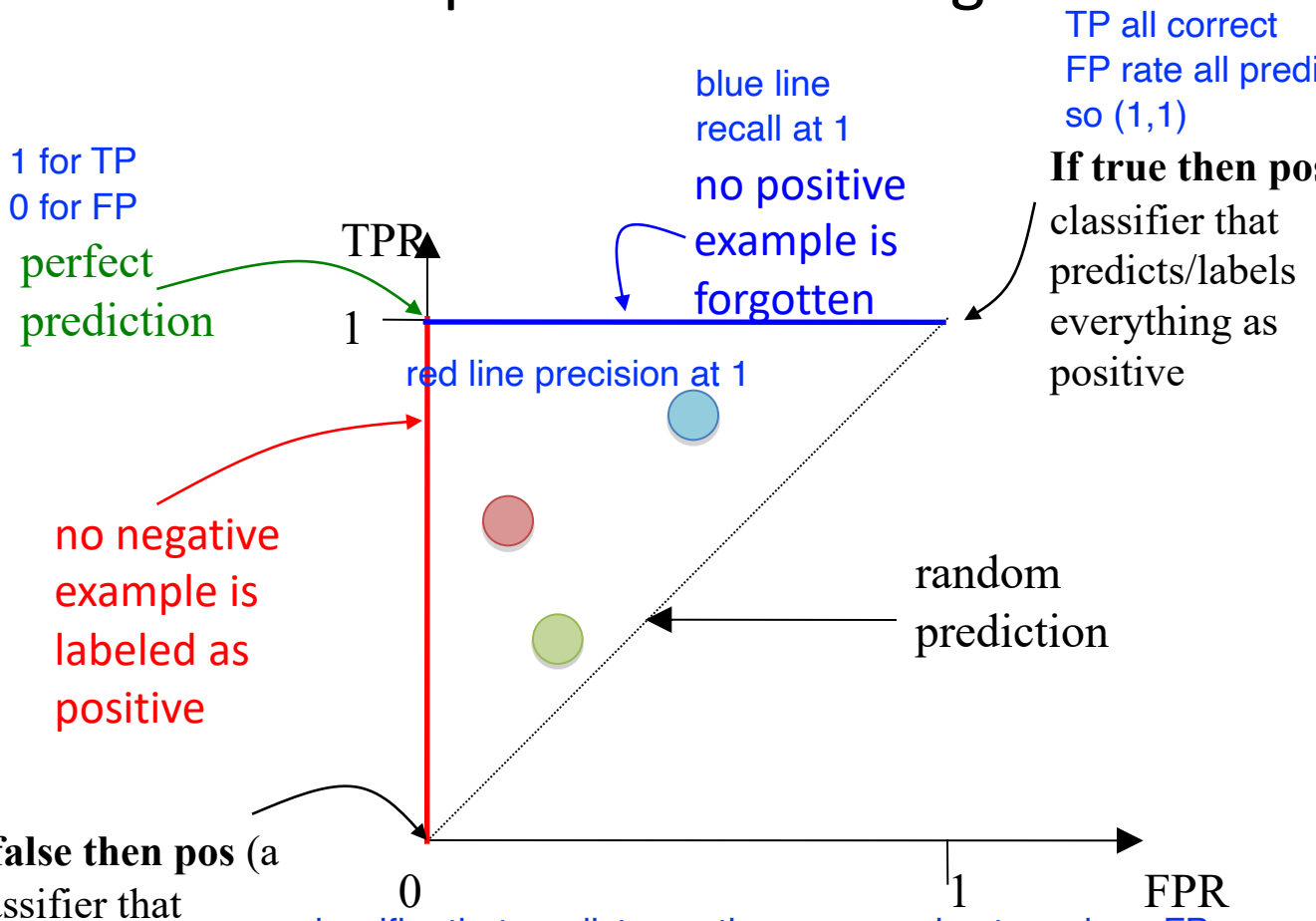
- ROC = "Receiver operating characteristic”
- ROC diagram plots *TPR vs FPR* of a classifier.

axis can be any 2 independenant vaues.
eg instead FPR, can have TNR
but cannot have TPR and FNR as they are functions of eachother

- Each point on the plot is a classifier.
- Allows to see how well a classifier performs:
  - given certain threshold (an internal parameter of a classifier);
  - given certain misclassification costs; and
  - given certain class distribution.

Classifier in ROC diagram

1 classifier = 1 point on ROC diagram



If true then pos (a classifier that predicts/labels everything as positive

True	Predicted	
	pos	neg
+	80	20
-	50	50

True	Predicted	
	pos	neg
+	60	40
-	20	80

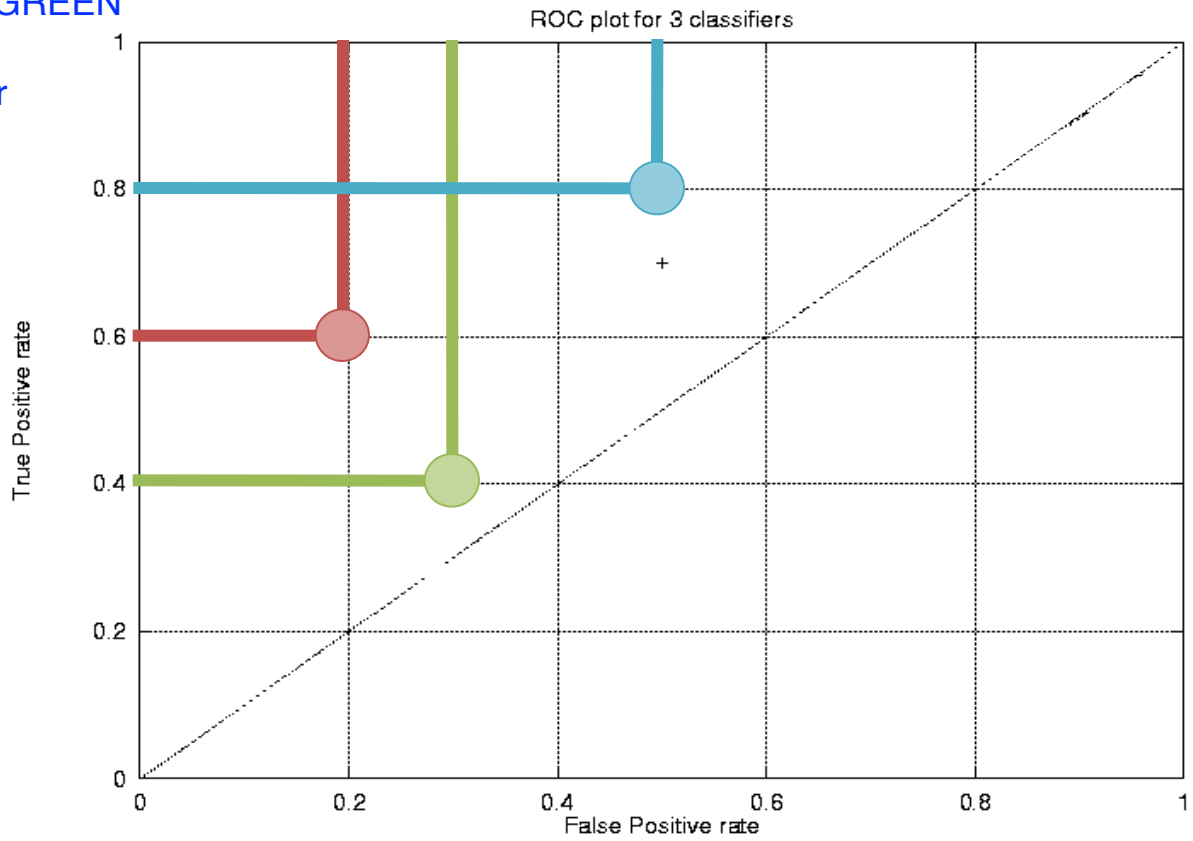
True	Predicted	
	pos	neg
+	40	60
-	30	70

if false then pos (a classifier that predicts/labels everything as negative

Dominance in the ROC Space

red dominATES GREEN

alike partial order

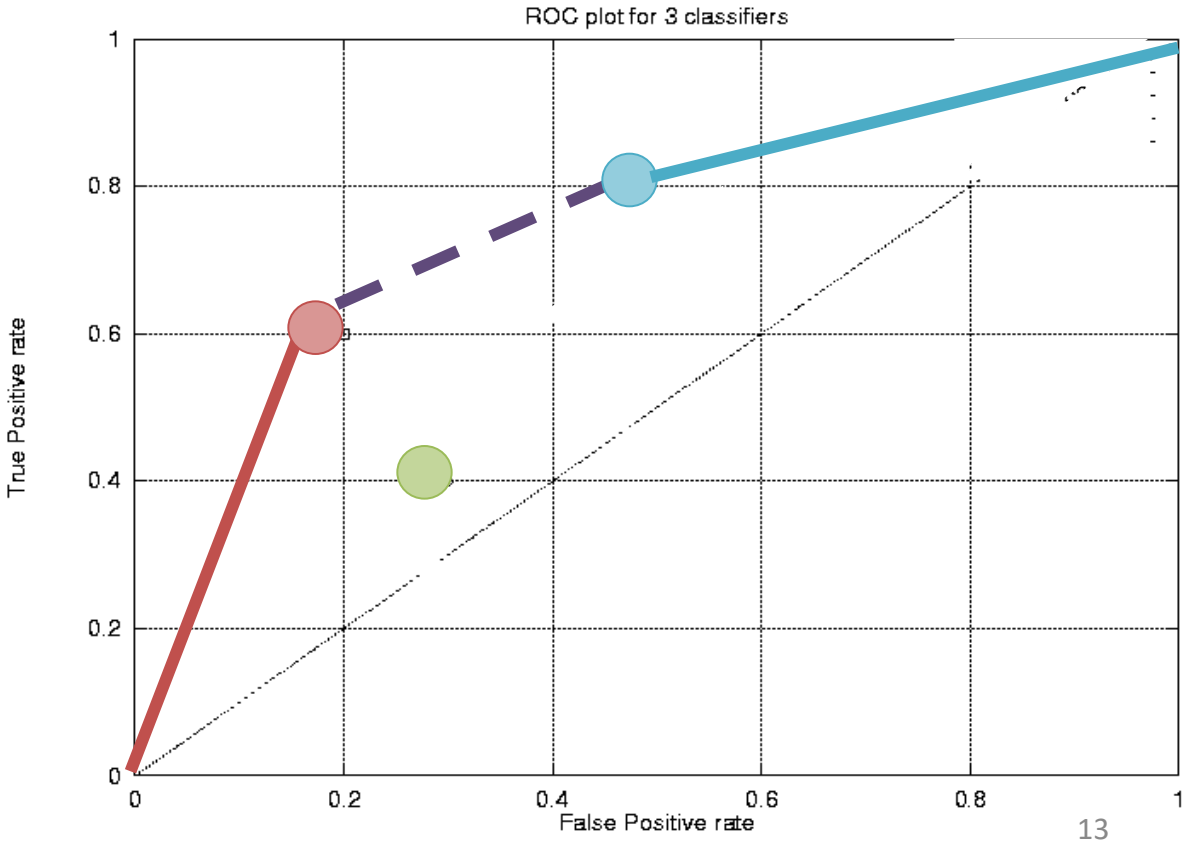


Classifier A dominates classifier B if and only if $TPR_A > TPR_B$ and $FPR_A < FPR_B$.

ROC Convex Hull (ROCCH)

- Dominant (non-dominated) classifiers can be connected by a straight line.
- Classifiers below ROCCH are always sub-optimal. eg green
- Points below the diagonal can be moved to other side by swapping the outputs.

- Any point of the line segment connecting two classifiers can be achieved by randomly choosing between them;



ROC of a classifier

Some classifiers can output their certainty about a prediction.

chaging threshold to have dif models of classifiers

decison tree can use purity

Examples:

(1) Decision trees:

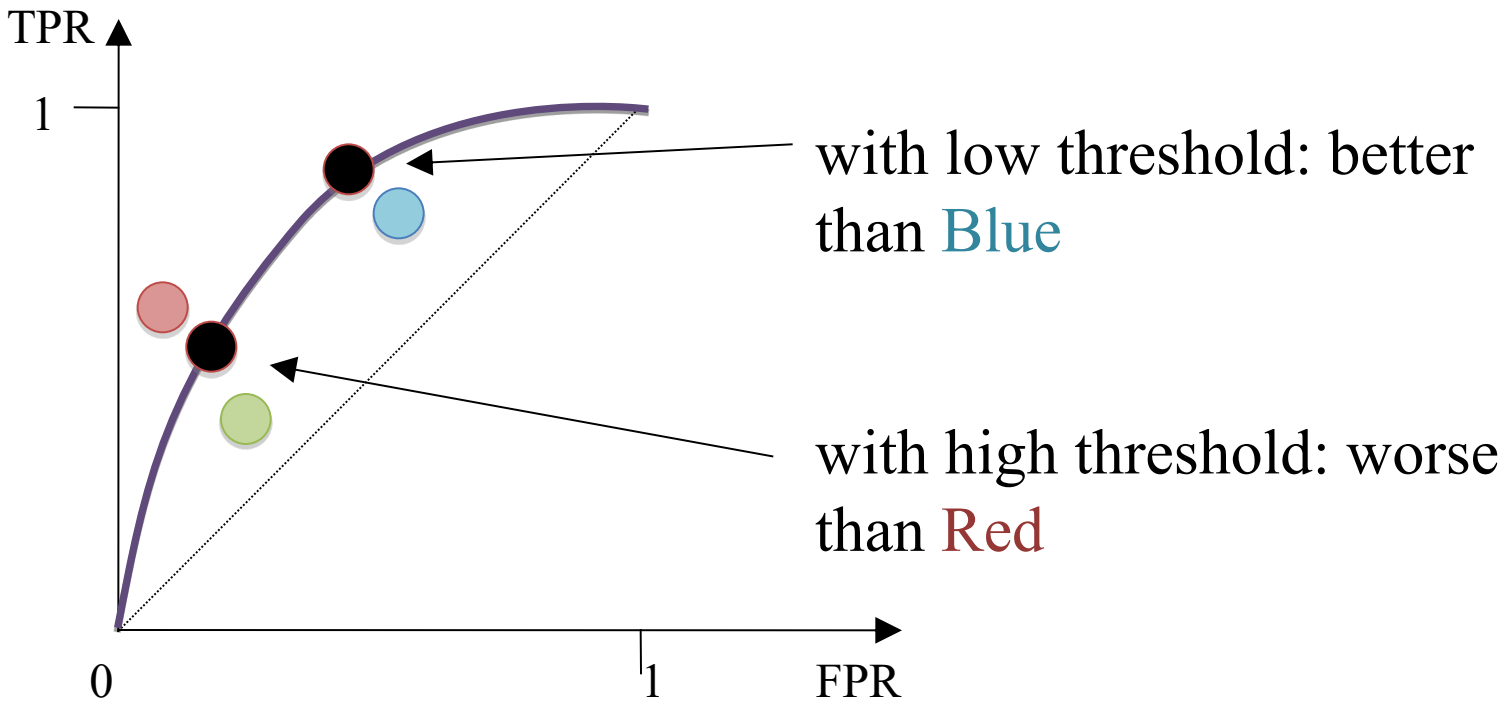
- use purity of the leaf used for prediction as a degree of certainty.
- e.g. leaf with 90% positives is more certain than leaf with 80% positives

(2) Logistic regression

- Simple way: if the single output < 0.5 then neg. otherwise pos.
- but 0.9 is more certainly positive than 0.51
- raise/lower threshold of 0.5: TP and FP go down or up

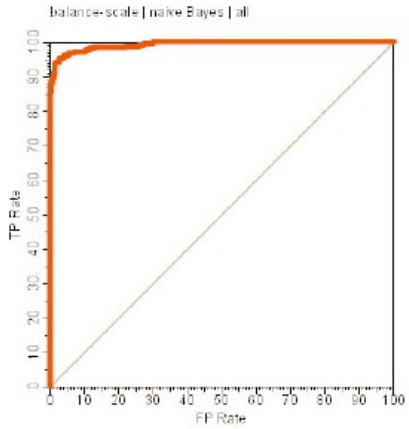
Each specific threshold yields one point on the ROC curve

changing threshold you can be on different parts on the curve
different classifiers



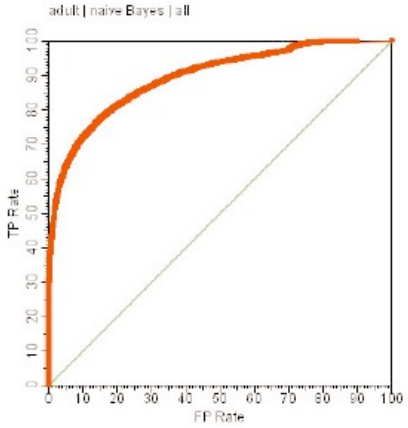
Example ROCs

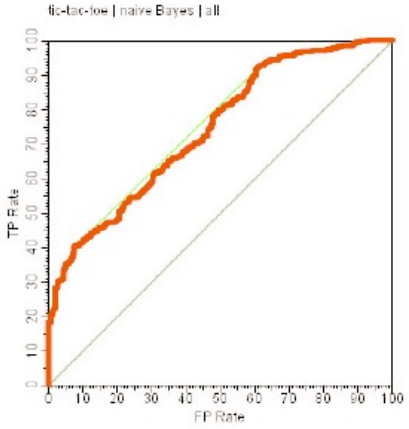
close to the top line for ideal predictions



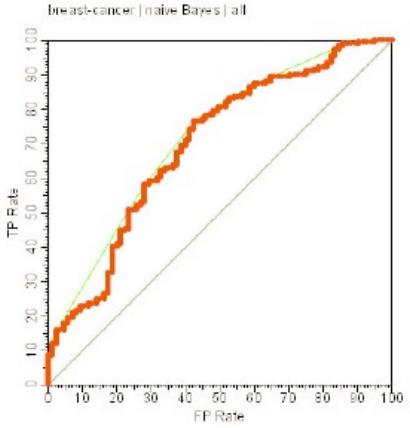
Good separation between the classes

Reasonable separation between the classes

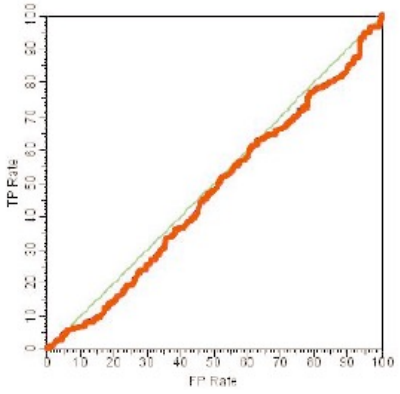




Fairly poor separation between the classes



Poor separation between the classes,

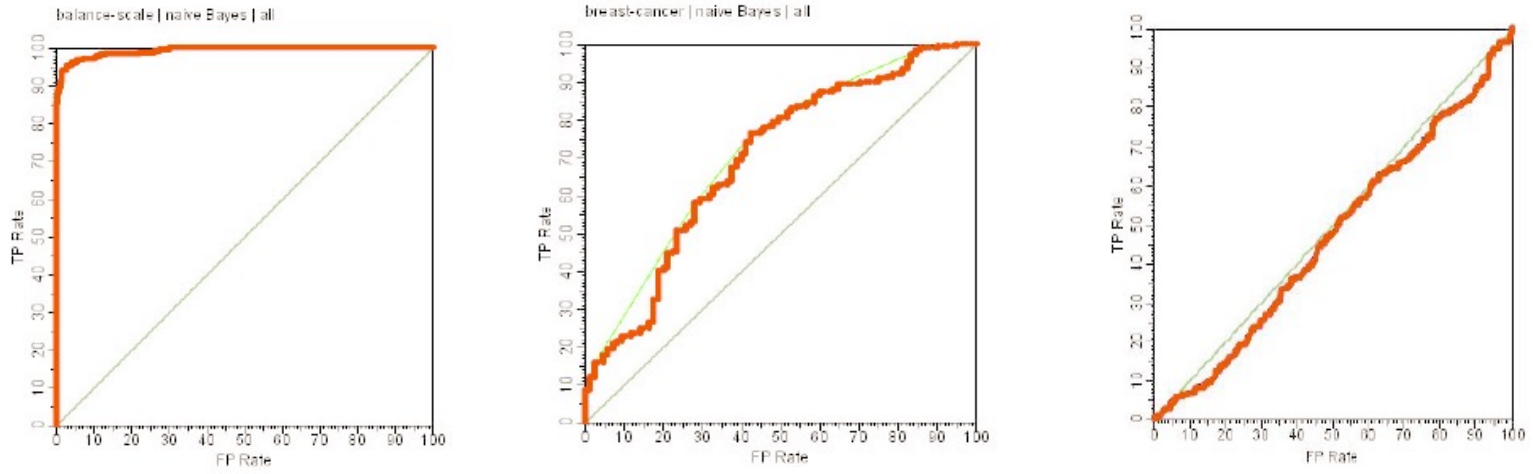


Random performance.

The AUC Metric

The area under ROC curve (AUC) assesses the performance in terms of separation of the classes.

ideal is an area of 1, random is 0.5, and others are 0,5 - 1



ROC curves can be used for internal optimization of classifiers (i.e. they are good objective functions when the actual costs are not known)

Costs in ROC diagram

Given misclassification costs:

- C_{FP}: cost of a false positive
- C_{FN}: cost of a false negative (undetected "+")

Average cost is

just changing to be 1-TPR,
reason is because using the same axis as ROC curve
otherwise same equation same meaning as prev slide^^^

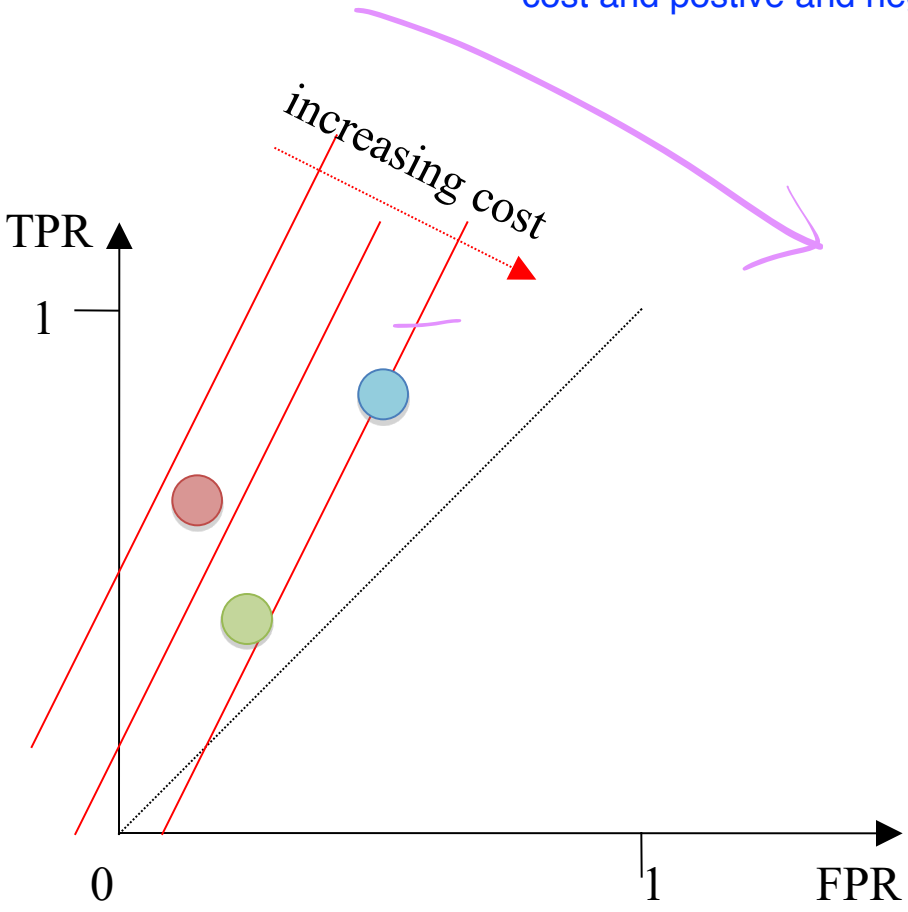
$$C = C_{FP} * FPR * N/(N+P) + C_{FN} * (1-TPR) * P/(P+N)$$

Lines of equal cost can be drawn in ROC diagram (straight lines)

contour
diff values of C, get different lines below graph

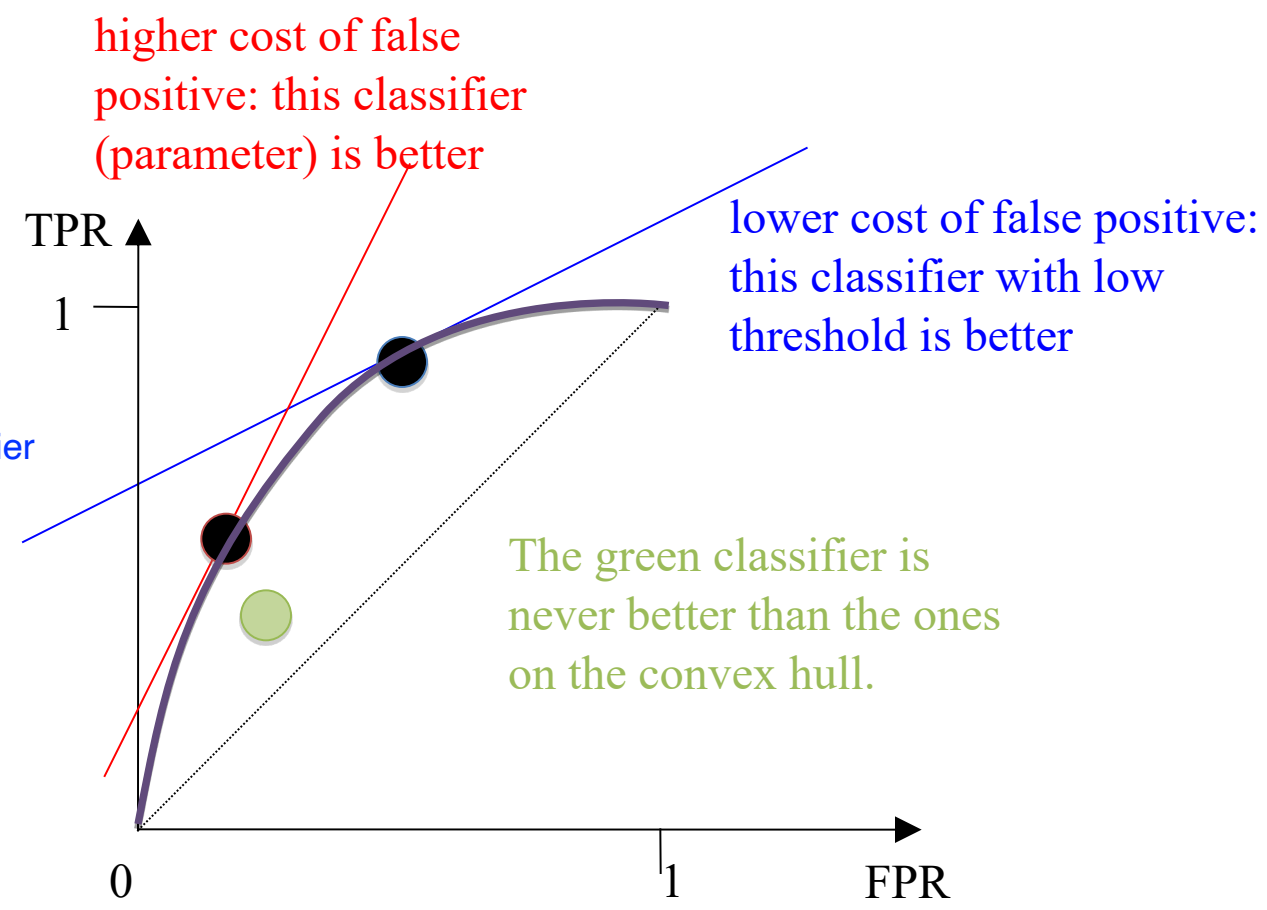
red with the smallest cost is best classifiers

slope of the line is from
cost and postive and negative examples



suppose a curve derieved from playing with the threshold

blue and red line are function of cost



given cost
try find best classifier
to meet conditions

Misclassification costs determine which classifier (or threshold) performs best.

References

- Bengio, Y., & Grandvalet, Y. (2005). Bias in estimating the variance of k-fold cross-validation. Statistical modeling and analysis for complex data problems, 75–95.
- Braga-Neto, U. M. (2004). Is cross-validation valid for small-sample microarray classification Bioinformatics, 20(3), 374–380. doi:10.1093/bioinformatics/btg419
- Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Statistics in Medicine, 26(29), 5320–5334.
- Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis, 53(11), 3735–3745. doi:10.1016/j.csda.2009.04.009
- Molinaro, A. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301–3307. doi:10.1093/bioinformatics/bti499
- Sauerbrei, W., & Schumacher1, M. (2000). Bootstrap and Cross-Validation to Assess Complexity of Data-Driven Regression Models. Medical Data Analysis, 26–28.
- Tibshirani, RJ, & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. Arxiv preprint arXiv:0908.2904.

References (2)

- Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
- Fawcett, Tom. "An introduction to ROC analysis." *Pattern recognition letters* 27.8 (2006): 861-874.
- Baker, Stuart G. "The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer." *Journal of the National Cancer Institute* 95.7 (2003): 511-515.