Metrics

20 January 2026

Alex Lyman

Performance Measures

Performance Measures

- Most common measure is accuracy
  - Summed squared error
  - Mean squared error
  - Classification accuracy

Issues with Accuracy

- Is 99% accuracy good; Is 30% accuracy bad?
  - Depends on baseline and problem complexity
- Error reduction (1-accuracy)
  - Absolute vs relative
  - 99.90% accuracy to 99.99% accuracy is a 90% relative reduction in error, but absolute error is only reduced by .09%.
  - 50% accuracy to 75% accuracy is a 50% relative reduction in error and the absolute error reduction is 25%.
  - Which is better?
- **Above assumes equal cost for all errors**
  - **Often have different error costs – e.g. Heart attack or not**

Quick aside on Relative vs Absolute

- **“Our new model reduces error by 50%!”**
- **Scenario A:** Error went from **80%** → **40%**. (Huge improvement).
- **Scenario B:** Error went from **2%** → **1%**. (Tiny improvement in raw numbers, but technically still 50%).
- The word "Percent" is relative.
- **Percentage Point (Absolute):** Simple Subtraction.
  - Old Error−New Error
  - *Example:* 10%−5%=5 points
- **Percent Change (Relative):** The Ratio.
  - Old / (Old−New)
  - *Example:* 10/(10−5)=0.5=50%

Binary Classification

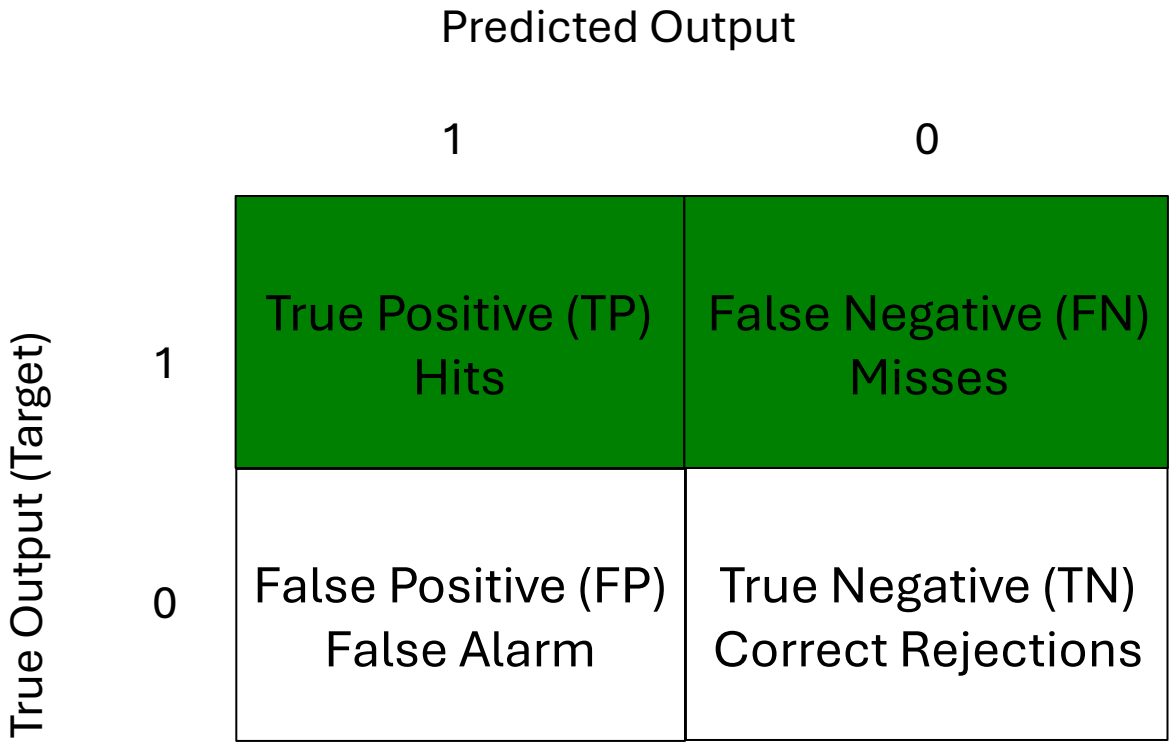
True Output (Target) \ Predicted Output	1	0
1	True Positive (TP) Hits	False Negative (FN) Misses
0	False Positive (FP) False Alarm	True Negative (TN) Correct Rejections

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

Recall



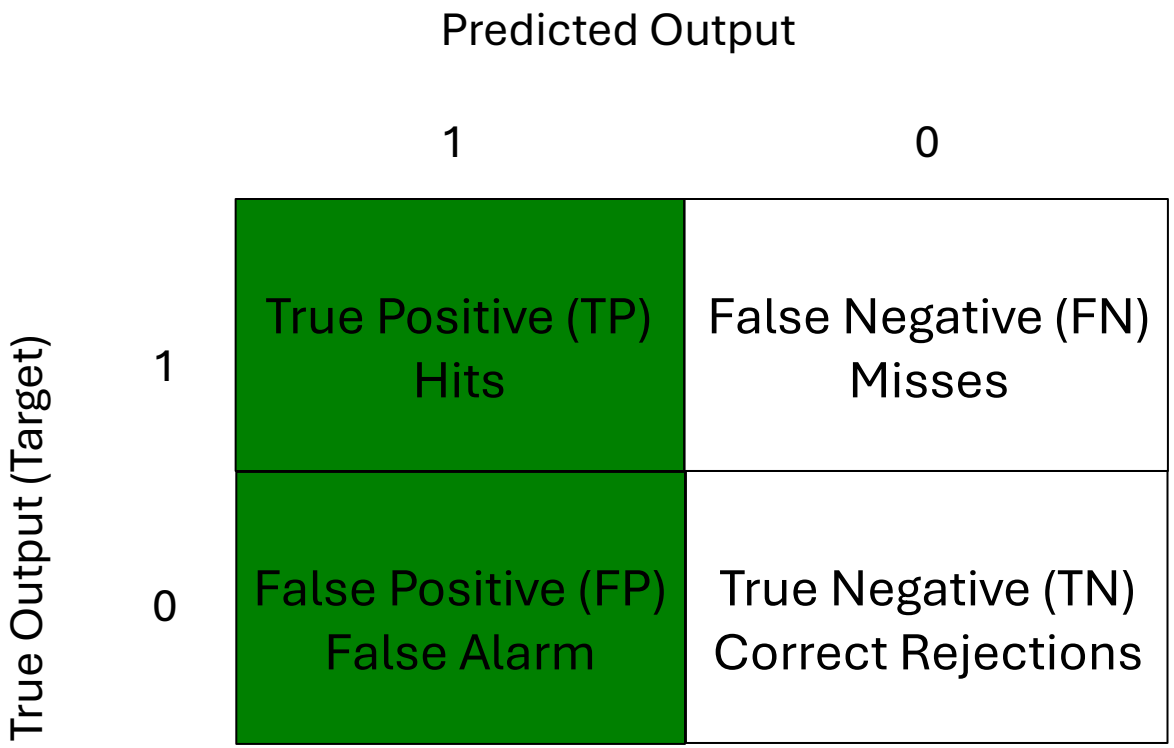
Recall = TP/(TP+FN)

The percentage of target true positives that were predicted as true positives, minimize false negatives

How to maximize?

Cancer Detection example

Precision



Precision = TP/(TP+FP)

The percentage of predicted true positives that are target true positives, minimize false positives

How to maximize?

Google Search Example

Other measures - Precision vs. Recall

- Find appropriate balance of Precision vs Recall for the task at hand, rather than just accuracy
- Can adjust ML parameters to accomplish the Precision vs Recall balance – Heart attack vs Google search
- Break even point: precision = recall
- F_1 or F-score = $2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$ - Harmonic average of precision and recall
- One especially useful situation is when there is highly skewed data output, where accuracy may be misleading

Bear Detector Example

- We go back to our fictional classifier where we want to detect dogs (vs bears). 97% of the pictures are dogs. 3% are bears. Our classifier just learns to classify everything as a dog and gets 97% accuracy.

$$\frac{TN + TP}{Total} = \frac{97 + 0}{100} = \mathbf{97\%}$$

- **Recall:** (How did we do on bears?)

$$\frac{TP}{TP + FN} = \frac{0}{0 + 3} = \mathbf{0\%}$$

- **Precision:** (When we predicted bear, were we right?)

$$\frac{TP}{TP + FP} = \frac{0}{0 + 0} = \mathbf{Undefined\ (or\ 0)}$$

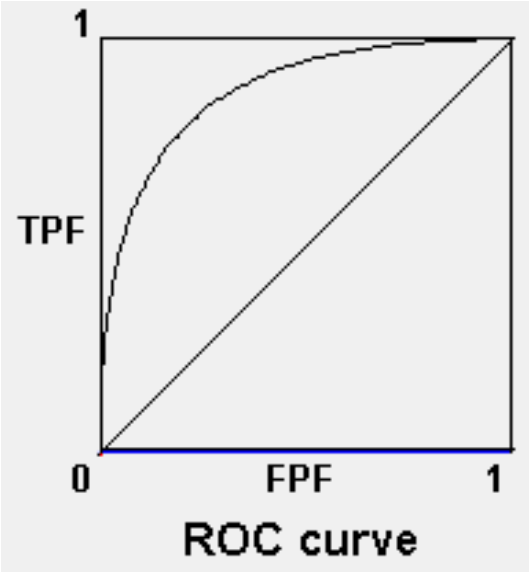
- **Harmonic Mean (F1-Score):** (The Balance)

$$2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \mathbf{0}$$

ROC Curves

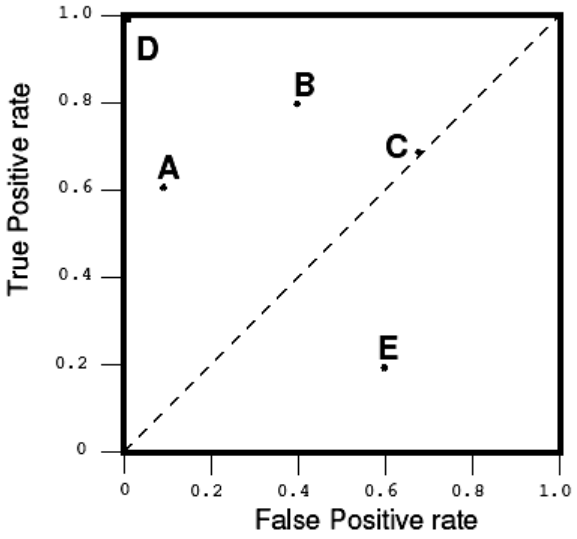
Receiver Operating Characteristic curve

- ROC curves were developed in the 1950's as a by-product of research into making sense of radio signals contaminated by noise. More recently it's become clear that they are remarkably useful in decision-making.
- True positive and False positive fractions are plotted as we move the dividing threshold. They look like:



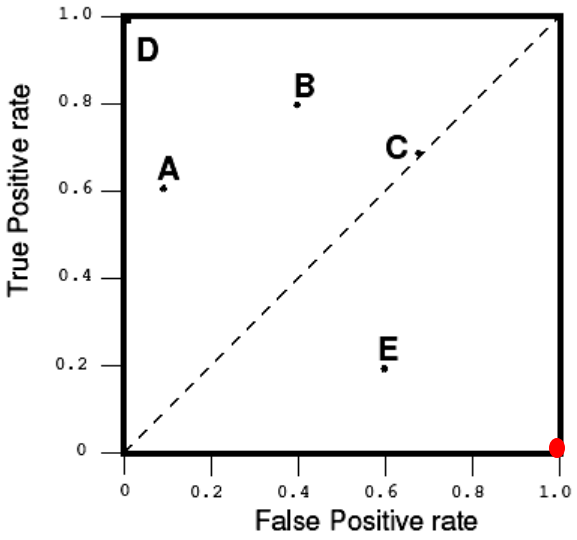
ROC Space

- ROC graphs are two-dimensional graphs in which TP rate is plotted on the Y axis and FP rate is plotted on the X axis.
- An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives).
- Figure shows an ROC graph with five classifiers labeled A through E.
- A discrete classier is one that outputs only a class label.
- Each discrete classier produces an (fp rate, tp rate) pair corresponding to a single point in ROC space.
- Classifiers in figure are all discrete classifiers.



Several Points in ROC Space

- **Lower left point (0, 0)** represents the strategy of never issuing a positive classification;
  - such a classier commits no false positive errors but also gains no true positives.
- **Upper right corner (1, 1)** represents the opposite strategy, of unconditionally issuing positive classifications.
- **Point (0, 1)** represents perfect classification.
  - D's performance is perfect as shown.
- Informally, one point in ROC space is better than another if it is to the northwest of the first
  - tp rate is higher, fp rate is lower, or both.



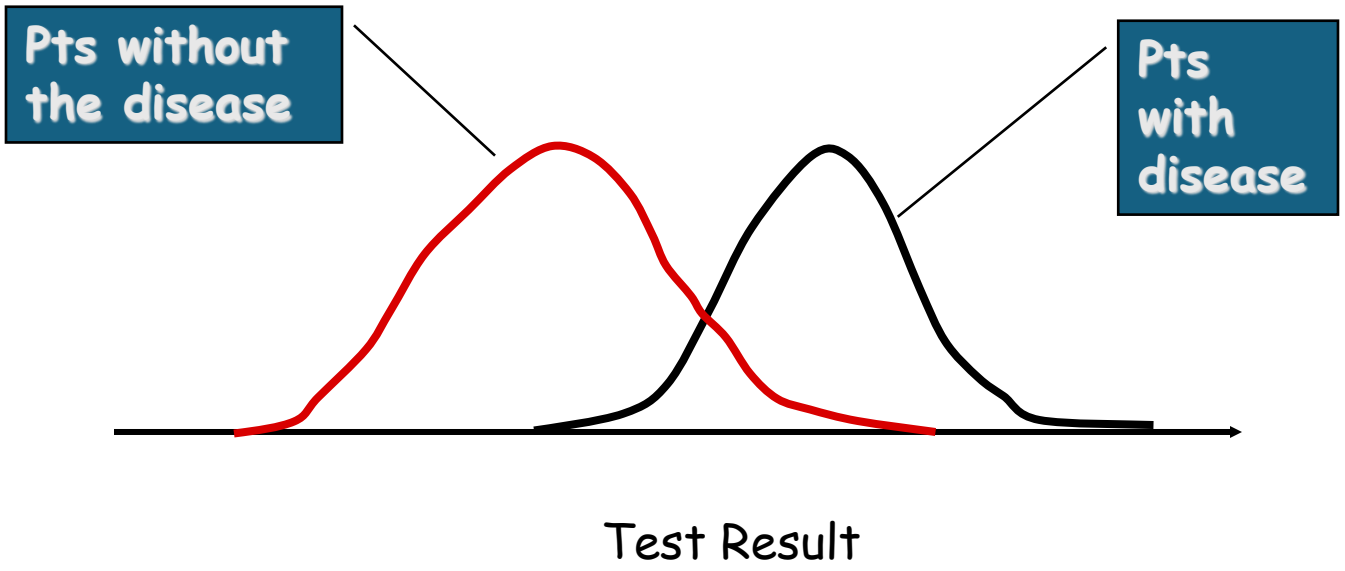
What's going on with this point?

ROC Curves and ROC Area

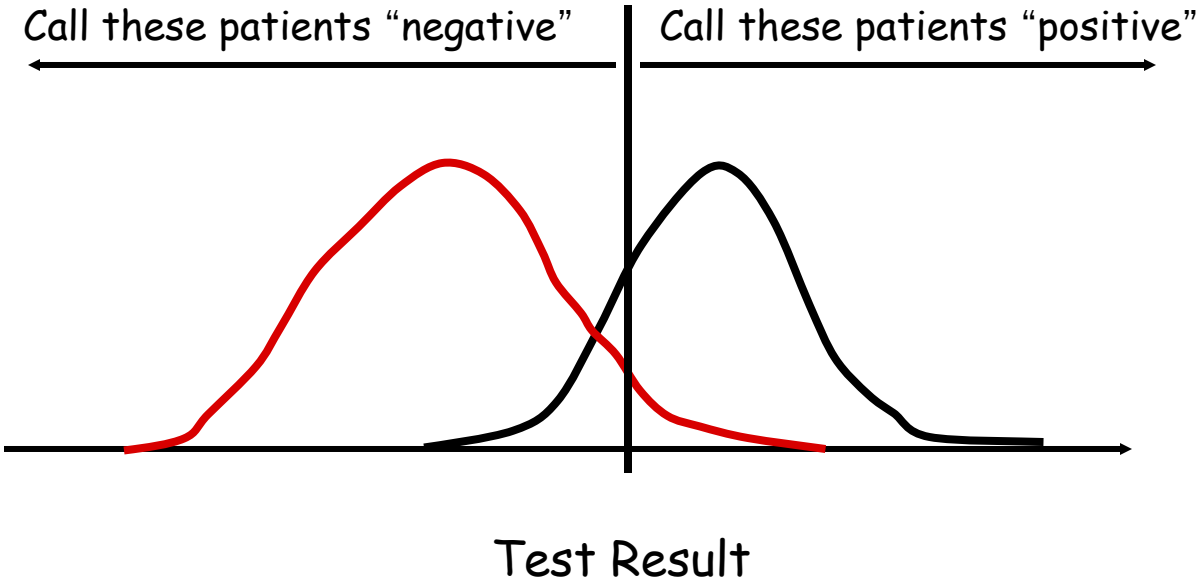
- Receiver Operating Characteristic
- Developed in WWII to statistically model false positive and false negative detections of radar operators
- Standard measure in medicine and biology
- True positive rate (sensitivity) vs false positive rate (1- specificity)
- True positive rate (Probability of predicting true when it is true)
  $P(Pred:T|T) = Sensitivity = Recall = TP/P = TP/(TP+FN)$
- False positive rate (Probability of predicting true when it is false)

  $P(Pred:T|F) = FP/N = FP/(TN+FP) = 1 – specificity (true negative rate) = 1 – TN/N = 1 - TN/(TN+FP)$
  - Want to maximize TPR and minimize FPR

Specific Example

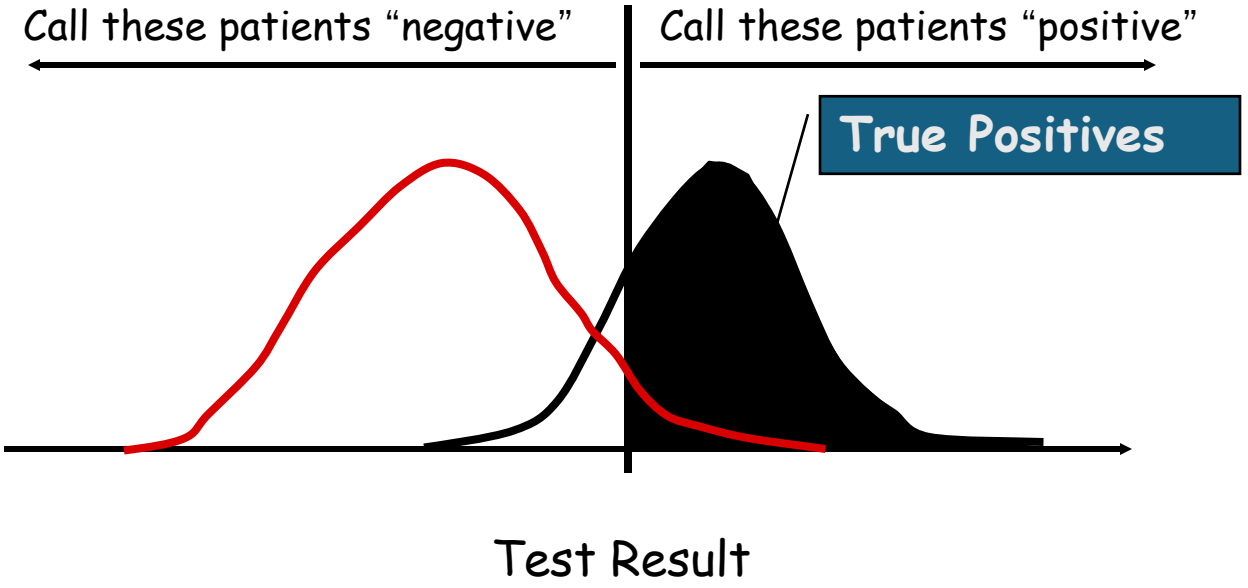


Threshold

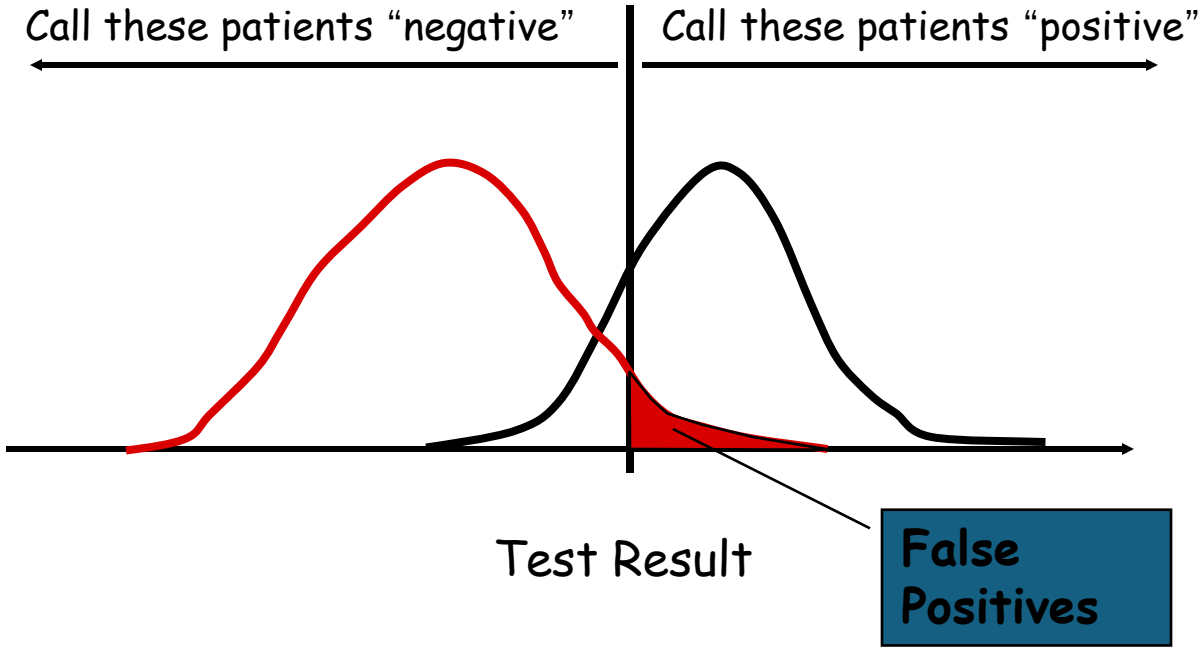


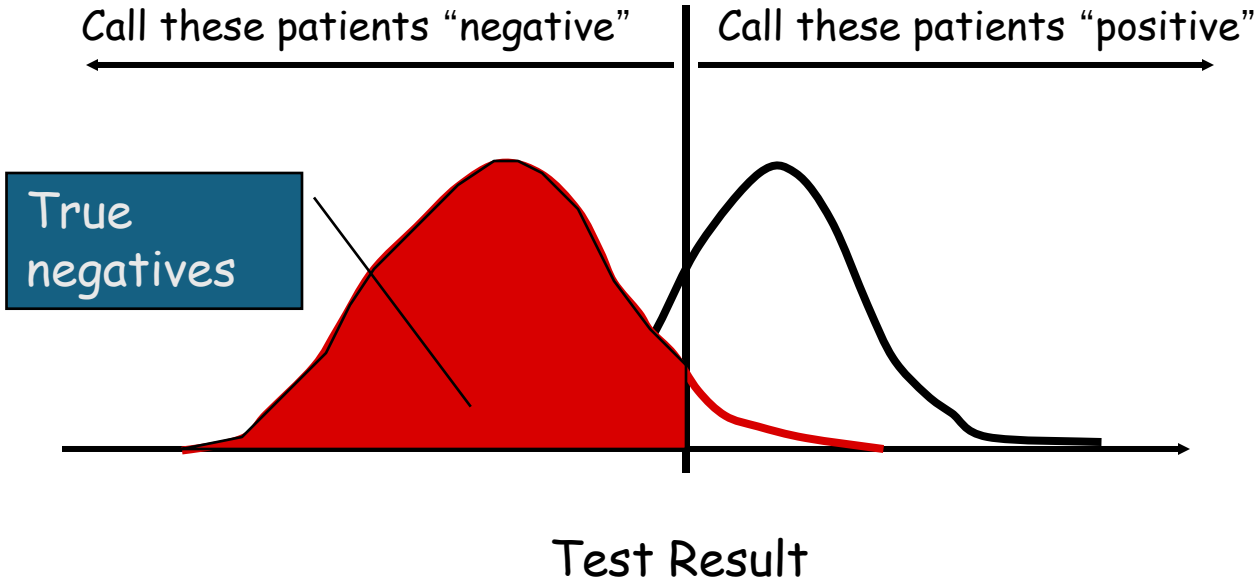
A classifier determines a line to separate the classes

Some definitions ...

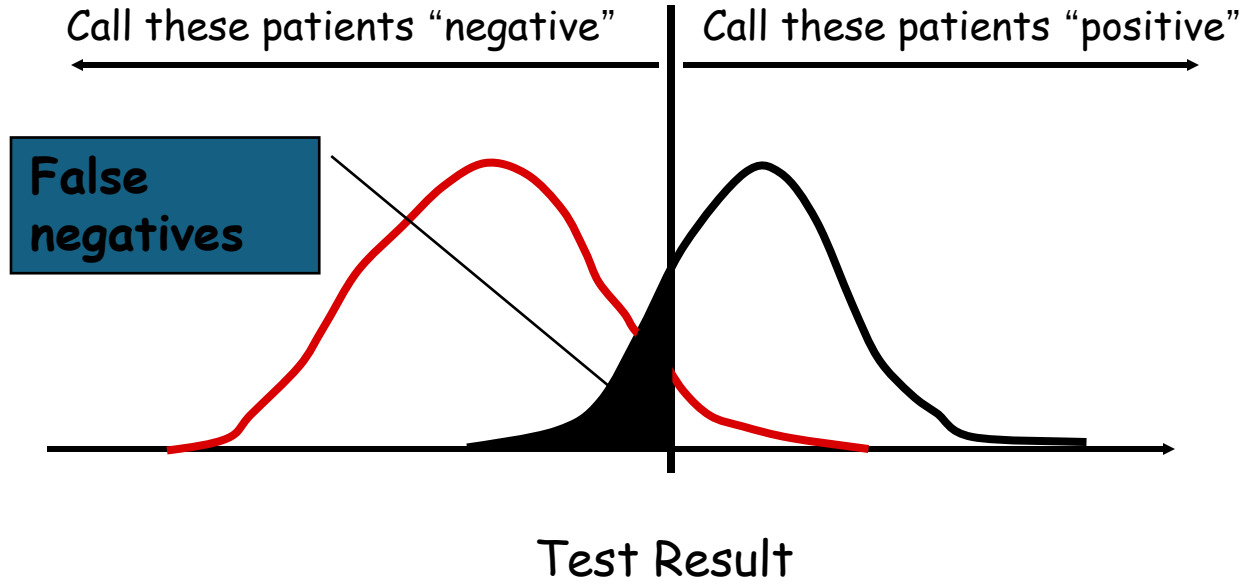


without the disease
with the disease


Call these patients "negative"
Call these patients "positive"
Test Result
False Positives
without the disease
with the disease

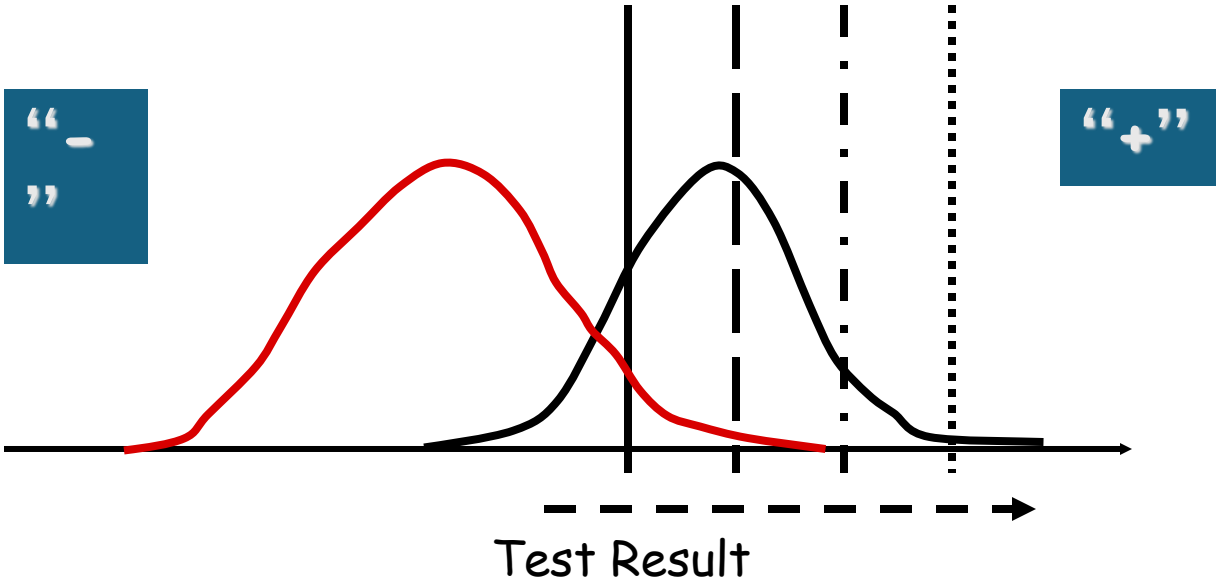


without the disease
with the disease



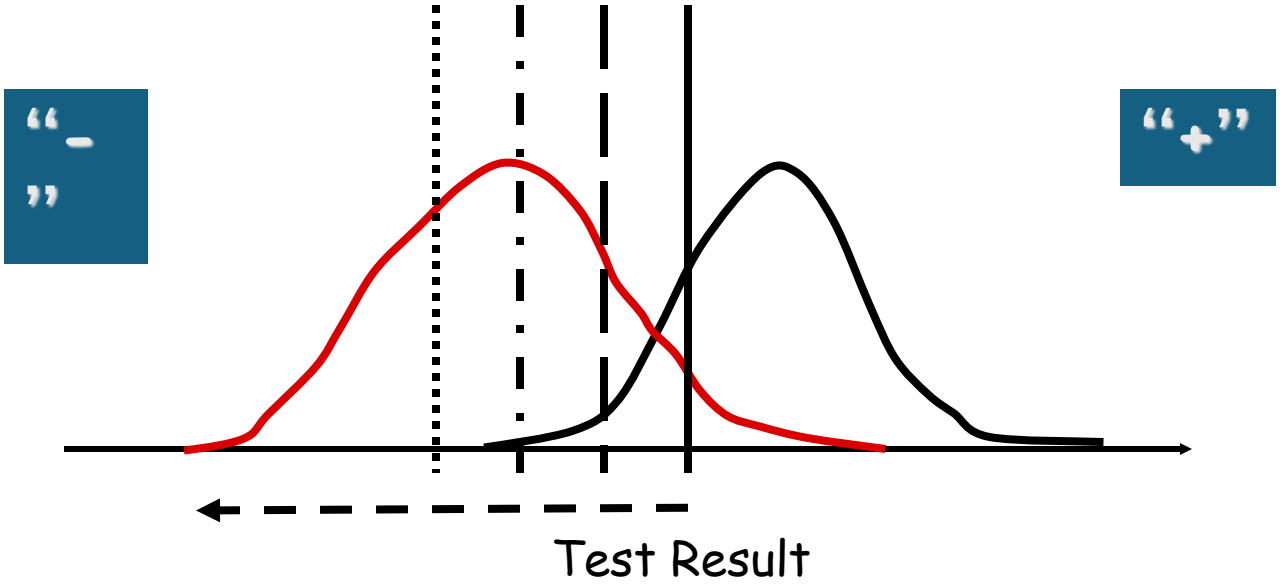
without the disease
with the disease

Moving the Threshold: right



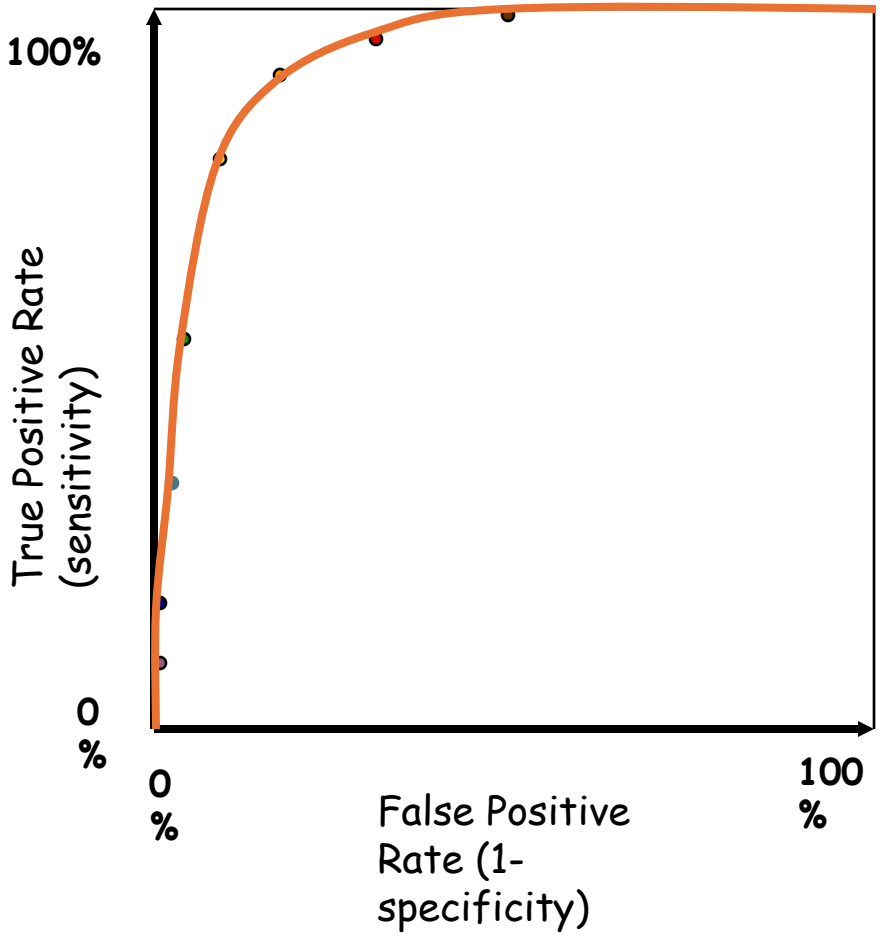
without the disease
with the disease

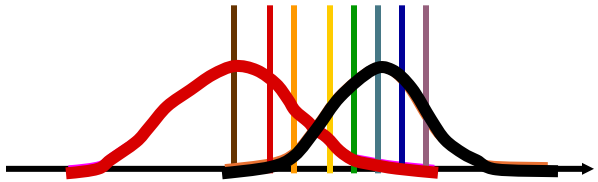
Moving the Threshold: left



without the disease
with the disease

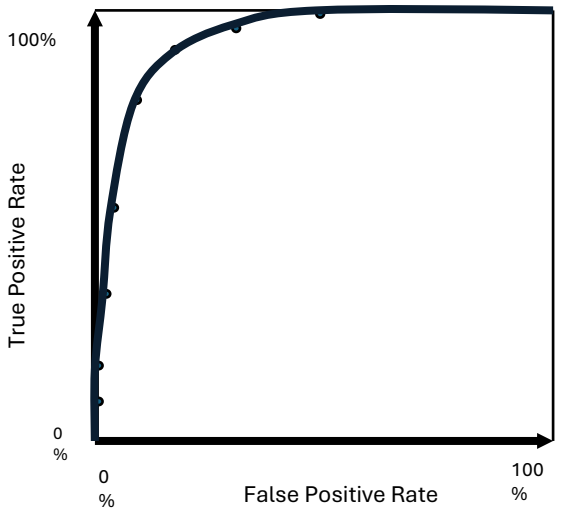
ROC curve


100%
True Positive Rate (sensitivity)
0 %
0 %
False Positive Rate (1-specificity)
100 %

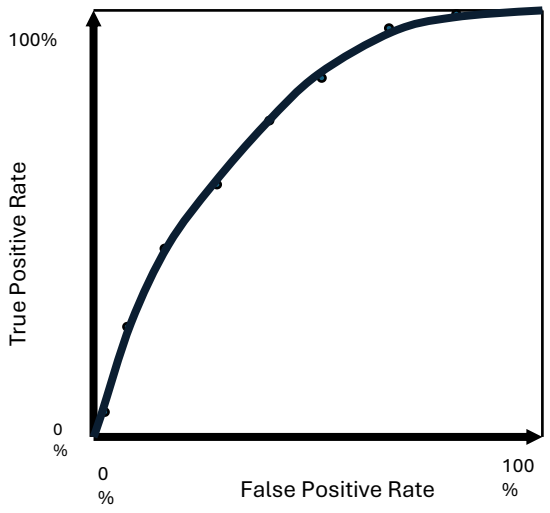


ROC curve comparison

A good test:

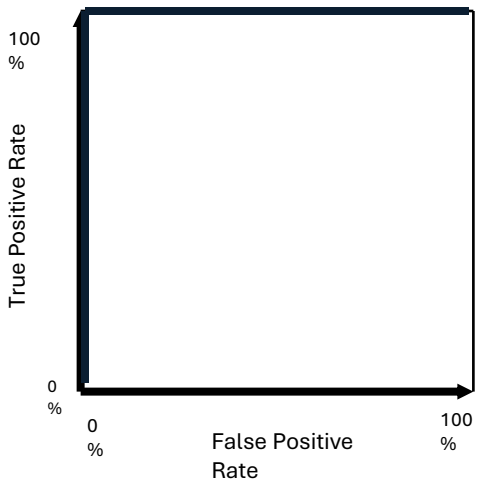


A poor test:



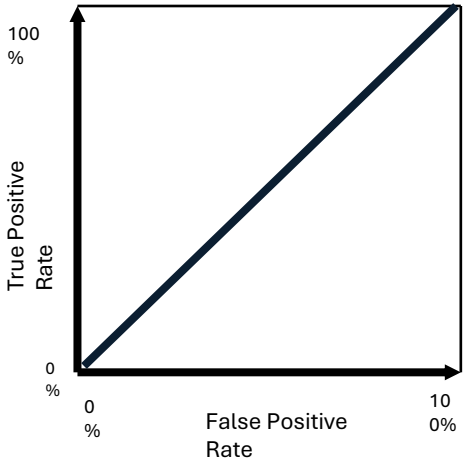
ROC curve extremes

Best Test:



The distributions don't overlap at all

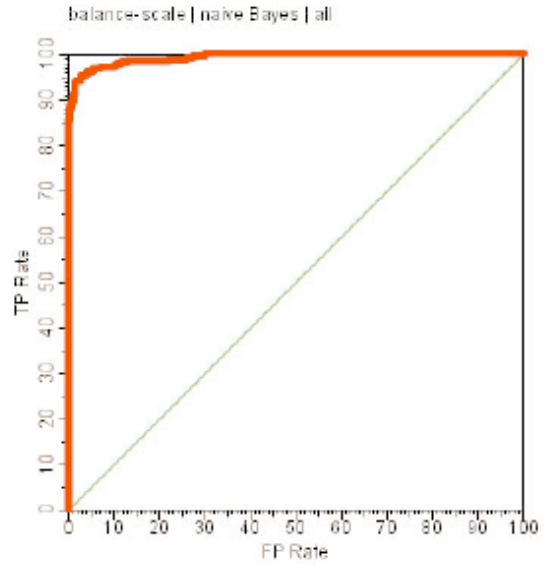
Worst test:



The distributions overlap completely

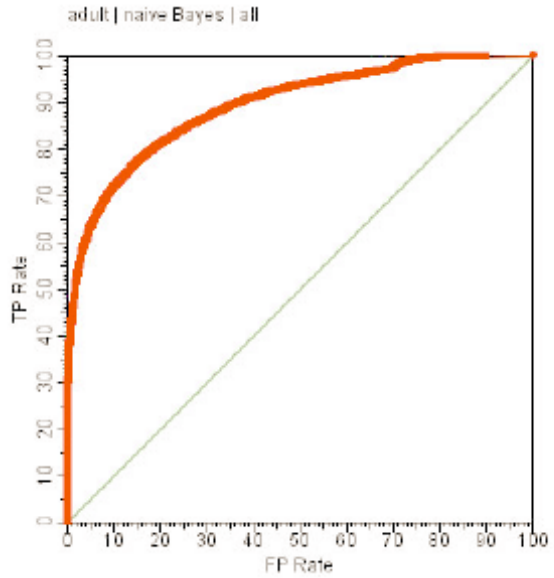
How Good are These?

ROC for one Classifier



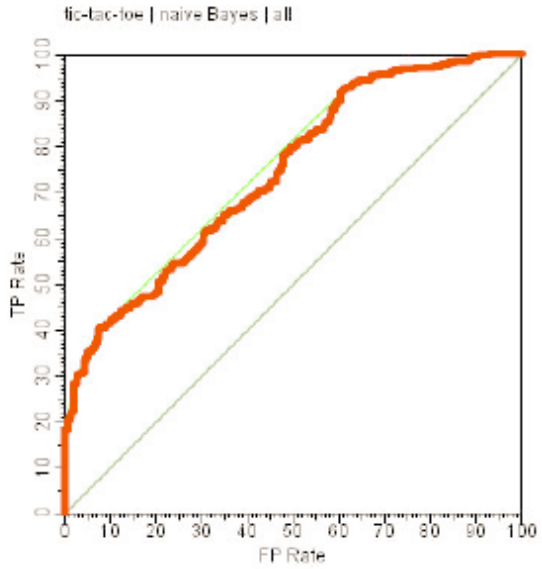
Good separation between the classes, convex curve.

ROC for one Classifier



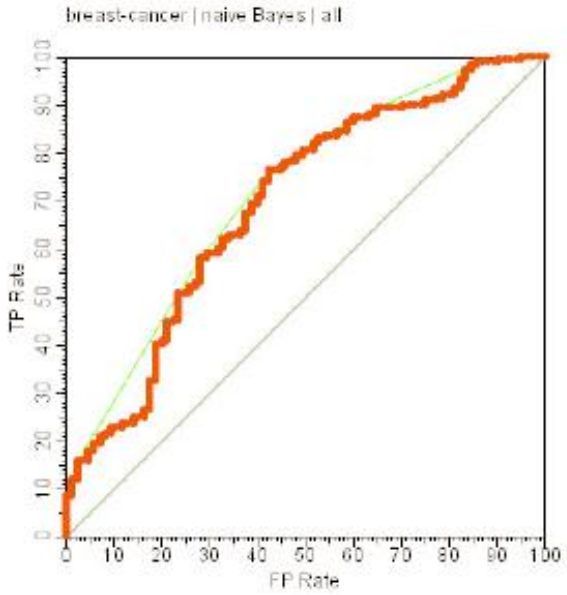
Reasonable separation between the classes, mostly convex.

ROC for one Classifier



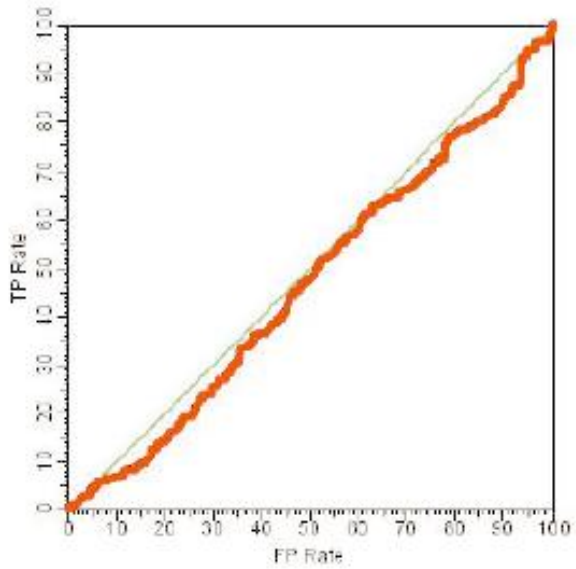
Fairly poor separation between the classes, mostly convex.

ROC for one Classifier



Poor separation between the classes, large and small concavities.

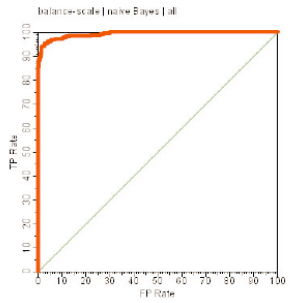
ROC for one Classifier

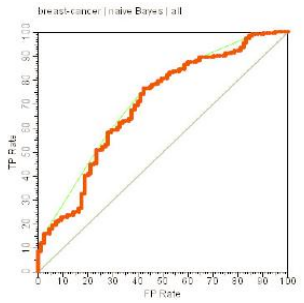


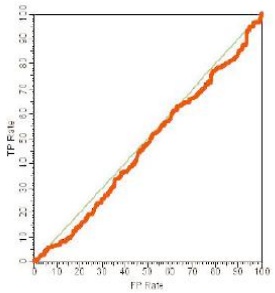
Random performance.

The AUC Metric

- The area under ROC curve (AUC) assesses the ranking in terms of separation of the classes.
- AUC estimates that randomly chosen positive instance will be ranked before randomly chosen negative instances.

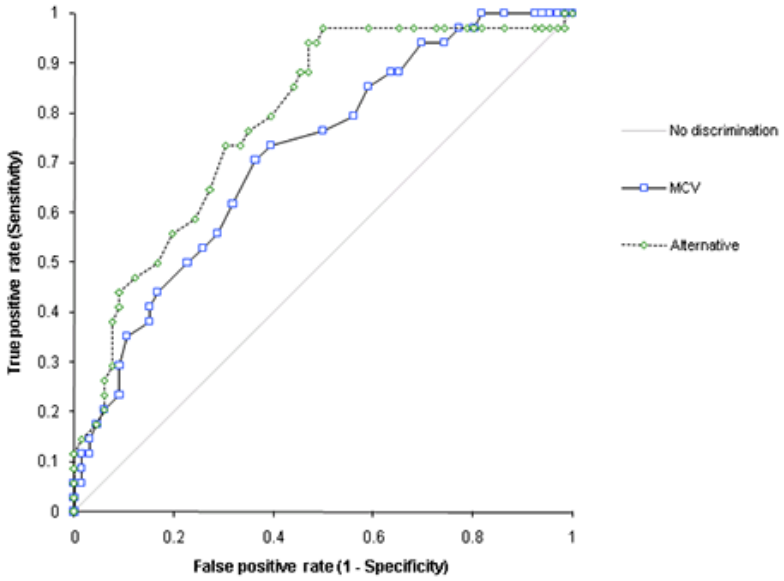




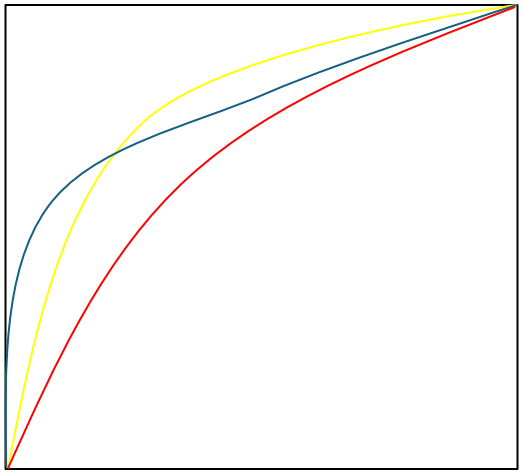


Comparing Models

- Highest AUC wins
- But pay attention to 'Occam's Razor'
  - 'the best theory is the smallest one that describes all the facts'
  - Also known as the 'parsimony principle'
  - If two models are similar, pick the simpler one



ROC Properties



- Area Properties
  - 1.0 - Perfect prediction
  - .9 - Excellent
  - .7 - Mediocre
  - .5 - Random
- ROC area represents performance over all possible cost ratios
- If two ROC curves do not intersect then one method dominates over the other
- If they do intersect then one method is better for some cost ratios, and is worse for others
  - Blue alg better for precision, yellow alg for recall, red neither
- Can choose method and balance based on goals

Performance Measurement Summary

- Other measures (e.g. Precision vs Recall, ROC, F-score) gaining popularity
- There are extensions to multi-output cases
  - However, medicine, finance, etc. have lots of two class problems
- Accuracy handles multi-class outputs and is still the most common measure but often combined with other measures like those above