Performance Analysis of Learners

- By: **Kaveh Kavousi**
- Department of Bioinformatics
- IBB (Institute of Biochemistry and Biophysics)
- University of Tehran

Null Hypothesis

In a statistical test, sample results are compared to possible population conditions by way of two competing hypotheses:

- The *null hypothesis* is a neutral or "uninteresting" statement about a population, such as "no change" in the value of a parameter from a previous known value or "no difference" between two groups.

- The other, the *alternative* (or *research*) *hypothesis* is the "interesting" statement that the person performing the test would like to conclude if the data will allow it.

p-value

- We wish to test a null hypothesis against an alternative hypothesis using a dataset. The two hypotheses specify two statistical models for the process that produced the data. The alternative hypothesis is what we expect to be true if the null hypothesis is false.

- We cannot prove that the alternative hypothesis is true but we may be able to demonstrate that the alternative is much more plausible than the null hypothesis given the data.

- This demonstration is usually expressed in terms of a probability, *p*-value, quantifying the strength of the evidence against the null hypothesis in favor of the alternative.

- The P-value represents, for any one test (one marker), the probability of falsely rejecting the null hypothesis - that is, calling a difference real when it is not.

p-value

If this *p*-value is very small, usually less than or equal to a threshold value previously chosen called the ***significance level*** (traditionally 5% or 1%), it suggests that the observed data is inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the other hypothesis accepted as true.

$P > 0.10$	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
$0.05 < P < 0.10$	Weak evidence against the null hypothesis in favor of the alternative
$0.01 < P < 0.05$	Moderate evidence against the null hypothesis in favor of the alternative.
$0.001 < P < 0.01$	Strong evidence against the null hypothesis in favor of the alternative.
$P < 0.001$	Very strong evidence against the null hypothesis in favor of the alternative.

http://wiki.c2b2.columbia.edu/workbench/index.php/ANOVA

Type I and Type II Errors

In statistics,

- A **type I error** (or **error of the first kind**) is the incorrect rejection of a true null hypothesis (False Positive = FP). Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't.

a test shows a patient to have a disease they are tested for, when in fact the patient does not have the disease, or that a medical treatment should cure a disease, when in fact it doesn't.

- A **type II error** (or **error of the second kind**) is the failure to reject the alternative hypothesis. (False Negative = FN).

a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease; or a clinical trial of a medical treatment failing to show that the treatment works when really it does

Type I and Type II Errors (Cont.)

- The rate of the **type I error** is called the commonly called a "**false alarm**" or *size* of the test and denoted by the Greek letter α . It usually equals the significance level of a test. In the case of a simple null hypothesis α is the probability of a type I error. If the null hypothesis is composite, α is the maximum (supremum) of the possible probabilities of a **type I error**.

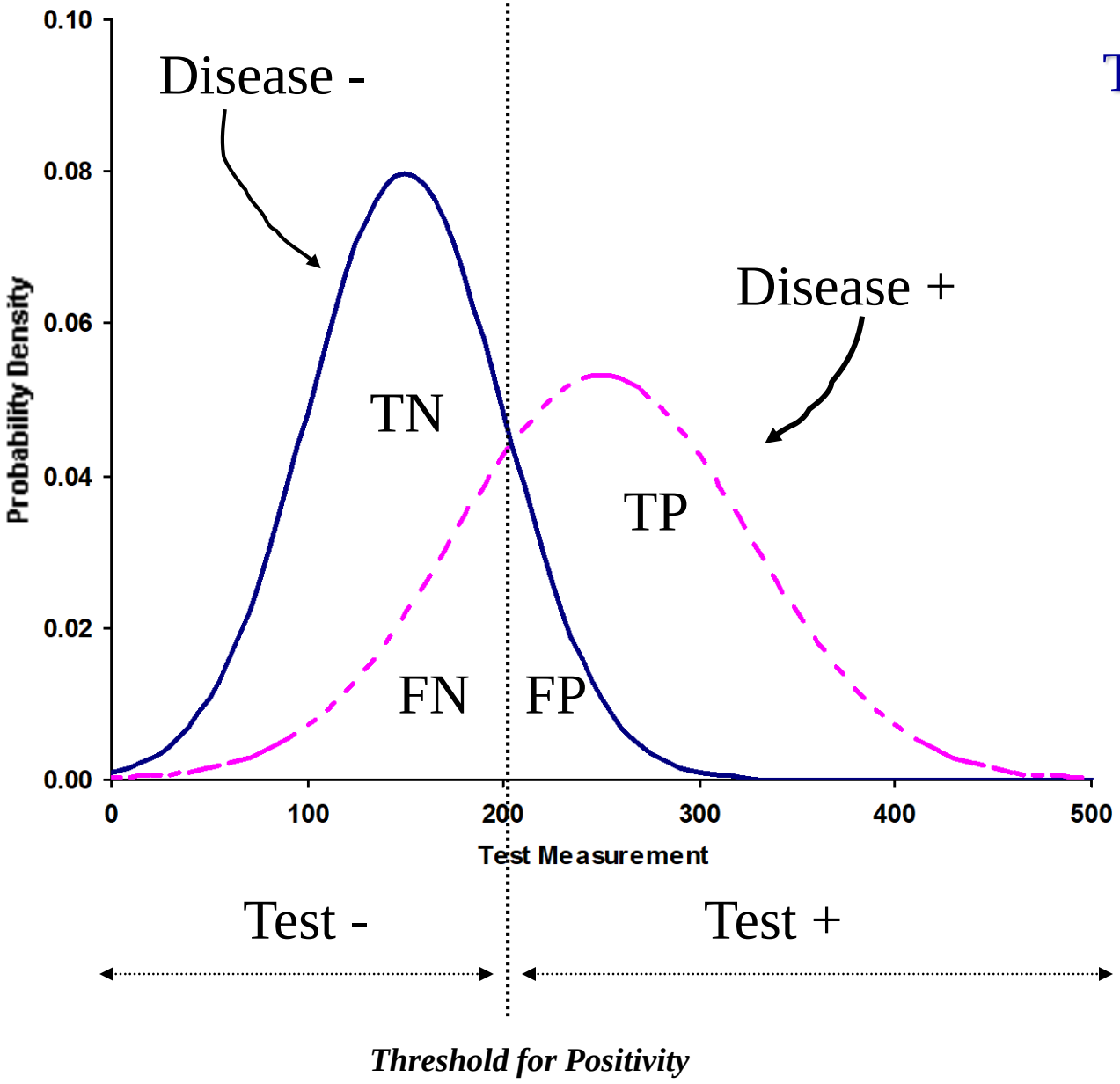
- The rate of the **type II error** is denoted by the Greek letter β and related to the power of a test (which equals 1-β).

- A statistical test can either reject (prove false) or fail to reject (fail to prove false) a null hypothesis, but never prove it true (i.e., failing to reject a null hypothesis does not prove it true).

Type I and Type II Errors (Cont.)

Confusion Matrix

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative



Specificity
TNF = TN/(TN+FP)
0.86
TNF + FPF = 1

Sensitivity
TPF = TP/(TP+FN)
0.73
TPF + FNF = 1

Test Characterization

- **SENSITIVITY** of a test is its ability to detect disease within a diseased population. It is calculated as the fraction of diseased patients correctly identified by the test. Also called the True Positive Fraction and TPF= TP/(TP+FN) where (TP + FN) is the number of patients with the disease.

  Can be thought of as the likelihood of spotting a positive case when presented with one. Or… the proportion of patients we find.

- **SPECIFICITY** of a test is its ability to identify the absence of disease in a disease free population. It is calculated as the fraction of non-diseased patients correctly identified by the test. Also called True Negative Fraction and TNF= TN/(TN + FP) where (TN + FP) is the number of patients that are disease free.

Test Characterization

- **ACCURACY** is the fraction of correct test results or diagnoses. It is calculated as the number of patients with correct test results divided by the whole patient population (TP+TN)/(TP+FP+TN+FN).

- **PREVALENCE (شيوع)** of the disease is calculated as the fraction of patients who have the disease (TP+FN)/(TP+FP+TN+FN).

Example: A study shows 90 true positives, 80 false positives, 20 true negatives and 10 false negatives. What are the sensitivity and specificity of the test?

Sensitivity = TPF

TPF = TP/(TP + FN)

TPF = 90/(90 +10) = 0.90

Sensitivity = 90%

Specificity = TNF

TNF = TN/(TN + FP)

TNF = 20/(20 +80) = 0.20

Specificity = 20%

Accuracy = ?

Evaluating Supervised Model Performance

The Confusion Matrix

- A matrix used to summarize the results of a supervised classification.
- Entries along the main diagonal are correct classifications.
- Entries other than those on the main diagonal are classification errors.

Table 2.5 • **A Three-Class Confusion Matrix**

		Computed Decision		
		c_1	c_2	c_3
True Classes	c_1	c_{11}	c_{12}	c_{13}
	c_2	c_{21}	c_{22}	c_{23}
	c_3	c_{31}	c_{32}	c_{33}

c_{11} is the number with true class “1” which are correctly classified as class “1”

c_{12} is the number with true class “1” which are mis-classified as class “2”

etc..

Two-Class Error Analysis

A Simple Confusion Matrix

		Computed Accept	**Computed Reject**
True	Accept	True Accept	False Reject
	Reject	False Accept	True Reject

PERFORMANCE MEASURES FOR CLASSIFICATION PROBLEMS

$$\textit{Average accuracy rate: } AA = \Pr\{\text{correct classification}\} = \frac{a+d}{a+b+c+d}$$

A Generic Confusion Matrix

		Predicted	
		Positive (P)	Negative (N)
Actual	Positive (P)	True Positive Cases (a)	False Negative Cases (b)
	Negative (N)	False Positive Cases (c)	True Negative Cases (d)

$$\textit{True positive rate: } TP = \Pr\{\text{predicted P}|\text{actually P}\} = \frac{a}{a+b}$$

$$\textit{True negative rate: } TN = \Pr\{\text{predicted N}|\text{actually N}\} = \frac{d}{c+d}$$

PERFORMANCE MEASURES FOR CLASSIFICATION PROBLEMS

A Generic Confusion Matrix

		Predicted	
		Positive (P)	Negative (N)
Actual	Positive (P)	True Positive Cases (a)	False Negative Cases (b)
	Negative (N)	False Positive Cases (c)	True Negative Cases (d)

$$\textit{Precision rate}: PR = \Pr\{\text{actually P}|\text{predicted P}\} = \frac{a}{a+c}$$

applying Bayes Theorem

$$PR = TP\,\frac{\Pr\{\text{actually P}\}}{\Pr\{\text{predicted P}\}}$$

PERFORMANCE MEASURES FOR CLASSIFICATION PROBLEMS

The Geometric Mean of TP and PR

$$\sqrt{TP \cdot PR}$$

It takes the value 1, when both TP and PR are equal to one; also the value 0, when either TP or PR equals zero. For all other values of TP and PR, their geometric mean $\sqrt{TP \cdot PR}$ is a number in the interval (0,1).

PERFORMANCE MEASURES FOR CLASSIFICATION PROBLEMS

The *F*-measure (Lewis and Gale 1994) is another performance measure where *TP* and *PR* rates may be combined. It is given by the equation:

$$F\text{-}Measure = \frac{(1+\beta)^2 \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision} \quad Or, \quad F = \frac{(\beta^2+1) * TP * PR}{\beta^2 * PR + TP},$$

Where β is a coefficient to adjust the relative importance of precision versus recall (usually, β =1)

where the β factor is a parameter that takes values from 0 to infinity and is used to control the influence of *TP* and *PR* separately. It can be shown that, when $\beta = 0$, then *F* reduces to *PR*, and conversely, when $\beta \to \infty$, then *F* approaches *TP*. Moreover, if for some data set $TP = PR$, then all four measures *F*, *TP*, *PR*, and $\sqrt{TP * PR}$ coincide. Finally, given that *TP* and *PR* are positive numbers less than 1, it can be shown that this is true for *F* as well.

The *F*-measure, is the harmonic mean of *precision* and *recall*. Gives one value, sort of an alternative to area under the ROC curve

PERFORMANCE MEASURES FOR CLASSIFICATION PROBLEMS

Another Definition for *F-measure*:

Hypothesized class	True class: p (positive)	True class: n (negative)
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$N = TN + FP; P = TP + FN$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

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

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

$$\text{sensitivity} = \text{recall}$$

$$\text{specificity} = \frac{\text{True negatives}}{\text{False positives} + \text{True negatives}} = 1 - fp\ rate$$

$$\text{positive predictive value} = \text{precision}$$

$$\text{F-measure} = \frac{2 \times precision \times recall}{precision + recall}$$

Matthews Correlation Coefficient

- The **Matthews correlation coefficient** is used in machine learning as a measure of the quality of binary (two-class) classifications, introduced by biochemist Brian W. Matthews in 1975. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

http://molbio.uoregon.edu/matthews/

- The MCC is a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

Matthews Correlation Coefficient

- While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best such measures.

- Other measures, such as the proportion of correct predictions (also termed **accuracy**), are not useful when the two classes are of very different sizes. For example, assigning every object to the larger set achieves a high proportion of correct predictions, but is not generally a useful classification.

Matthews Correlation Coefficient

- The MCC can be calculated directly from the confusion matrix using the formula:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one and hence MCC=0**

The original formula equal to above was:

$$\mathrm{N} = TN + TP + FN + FP$$

$$\mathrm{S} = \frac{TP + FN}{N}$$

$$\mathrm{P} = \frac{TP + FP}{N}$$

$$\mathrm{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$

$$\mathrm{MCC} = \sqrt{PPV \times TPR \times TNR \times NPV} - \sqrt{FDR \times FNR \times FPR \times FOR}$$

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405** (2): 442–451. doi:10.1016/0005-2795(75)90109-9

Advantages of MCC over accuracy and F1 score

- Although accuracy and F1 score are widely employed in statistics, both can be misleading, since they do not fully consider the size of the four classes of the confusion matrix in their final score computation.

- Suppose, you have a very imbalanced validation set made of 100 elements, 95 of which are positive elements, and only 5 are negative elements, and suppose also you made some mistakes in designing and training your machine learning classifier, and now you have an algorithm which always predicts positive. By applying your only-positive predictor to your imbalanced validation set, therefore, you obtain values for the confusion matrix categories: TP = 95, FP = 5; TN = 0, FN = 0.

- These values lead to the following performance scores: accuracy = 95%, and F1 score = 97.44%. By reading these over-optimistic scores, then you will be very happy and will think that your machine learning algorithm is doing an excellent job. Obviously, you would be on the wrong track.

Ref: "Ten quick tips for machine learning in computational biology" (BioData Mining, 2017)

Confusion Matrix or Contingency Table

		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population	
	Negative likelihood ratio (**LR−**) = FNR/TNR	False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR−				

Confusion Matrix or Contingency Table

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

---

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Miss Rate or False Negative Rate (FNR)

$$FNR = FN/(FN + TP)$$

accuracy (ACC)

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

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Informedness

$$TPR + SPC - 1$$

Markedness

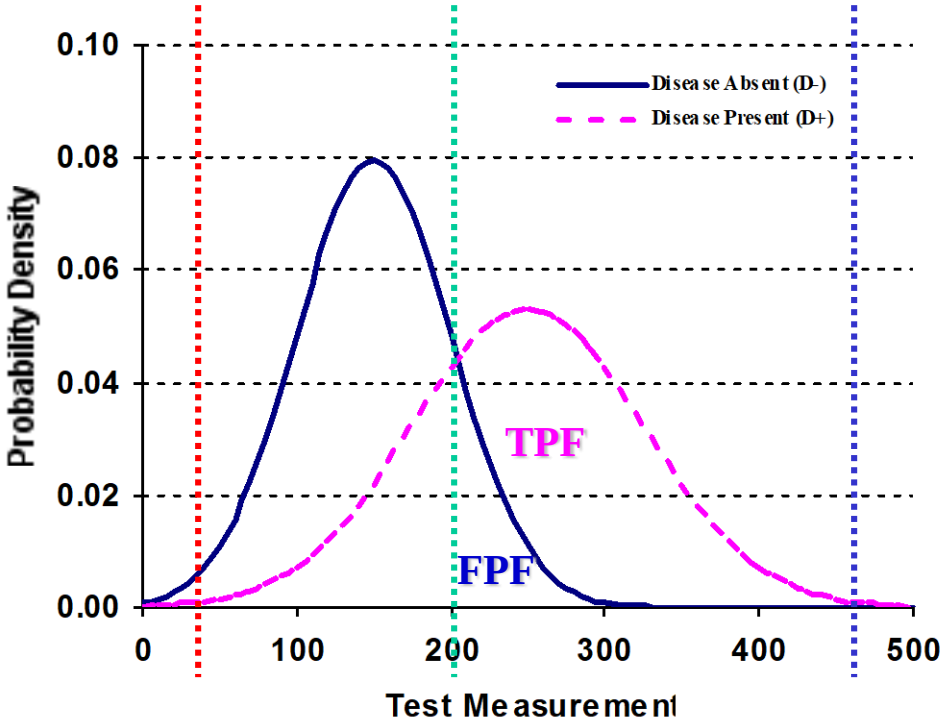
$$PPV + NPV - 1$$

Fawcelt, Tom (2006). "An Introduction to ROC Analysis". *Pattern Recognition Letters* **27** (8): 861 – 874. doi:10.1016/j.patrec.2005.10.010

ROC

Receiver Operating Characteristic- historic name from *radar* studies

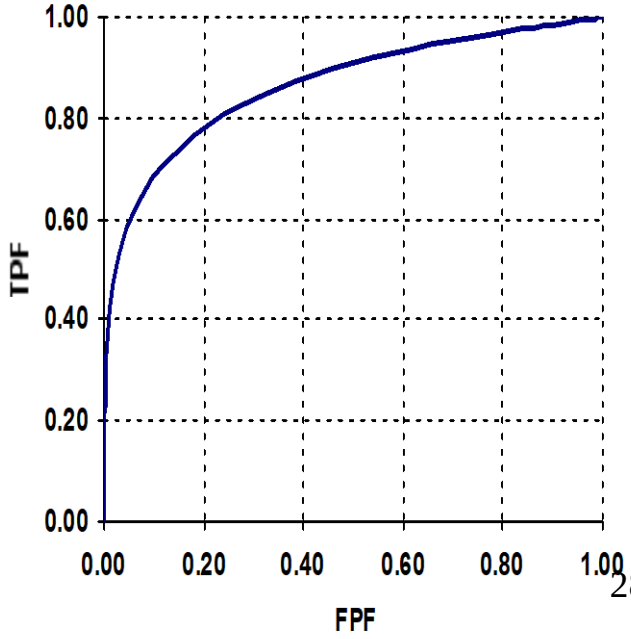
- The ROC curve is used to compare overall performance (sensitivity and specificity) of a test.
- An ROC curve is a graph of the True Positive Fraction (sensitivity) vs. False Positive Fraction (1-specificity) of a test as the threshold for positive result is changed.



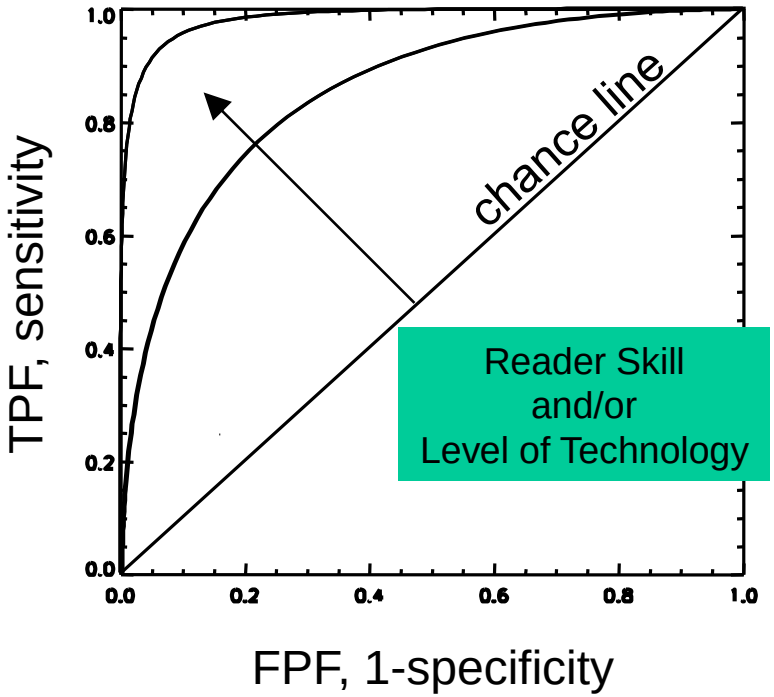
ROC Curve

D- mean = 150, SD = 50

D+ mean = 250, SD = 75



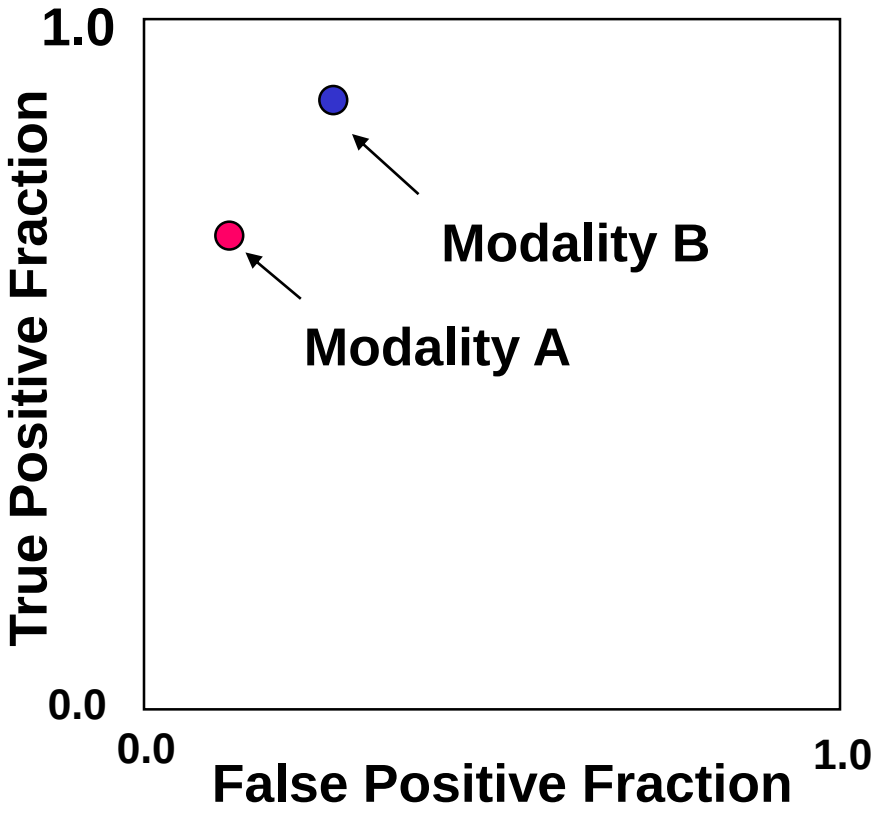
Entire ROC curve



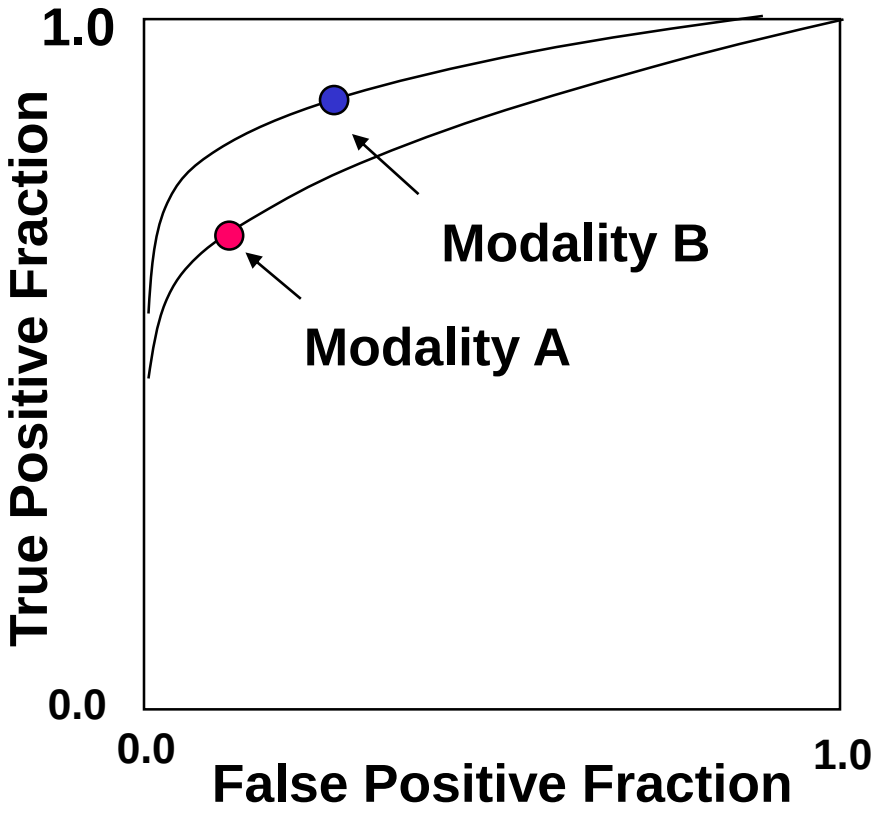
Quantifying ROC Curves

- The area under an ROC curve is a measure of overall performance.
- The maximum area is 1.0
- Useless test is the diagonal line from 0.0 to 1.0 and has area under ROC =0.5, so a more meaningful measure is the area in excess of 0.5.
- As test performance improves, the curve moves towards the upper left corner and the area under ROC increase.

Dilemma: Which modality is better?

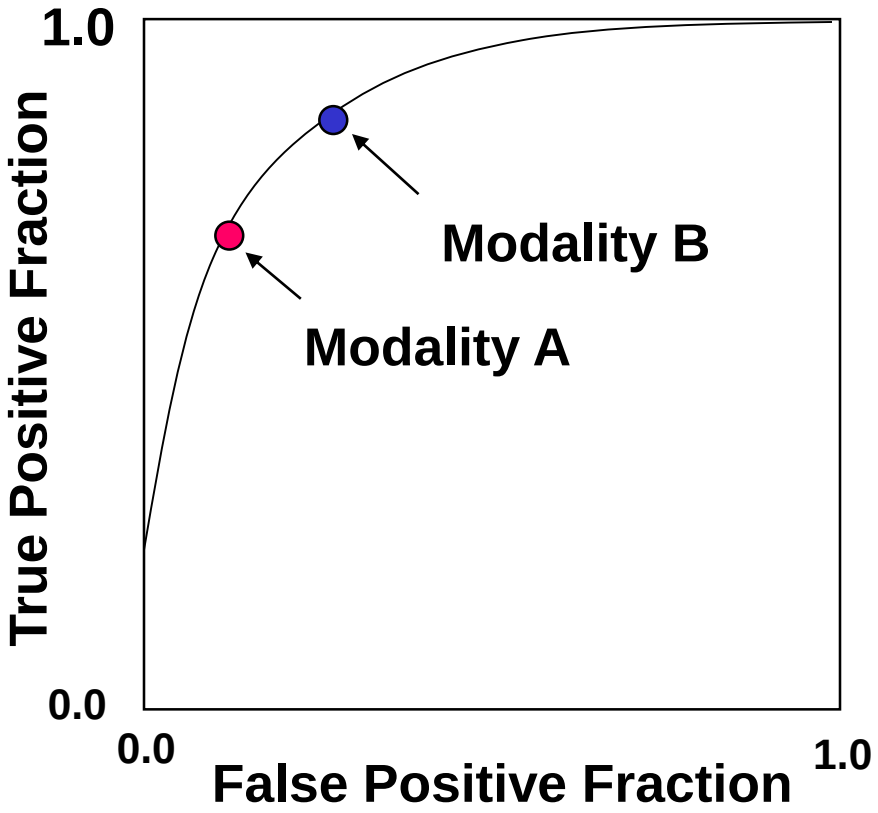


ROCs
(<u>one</u> outcome)



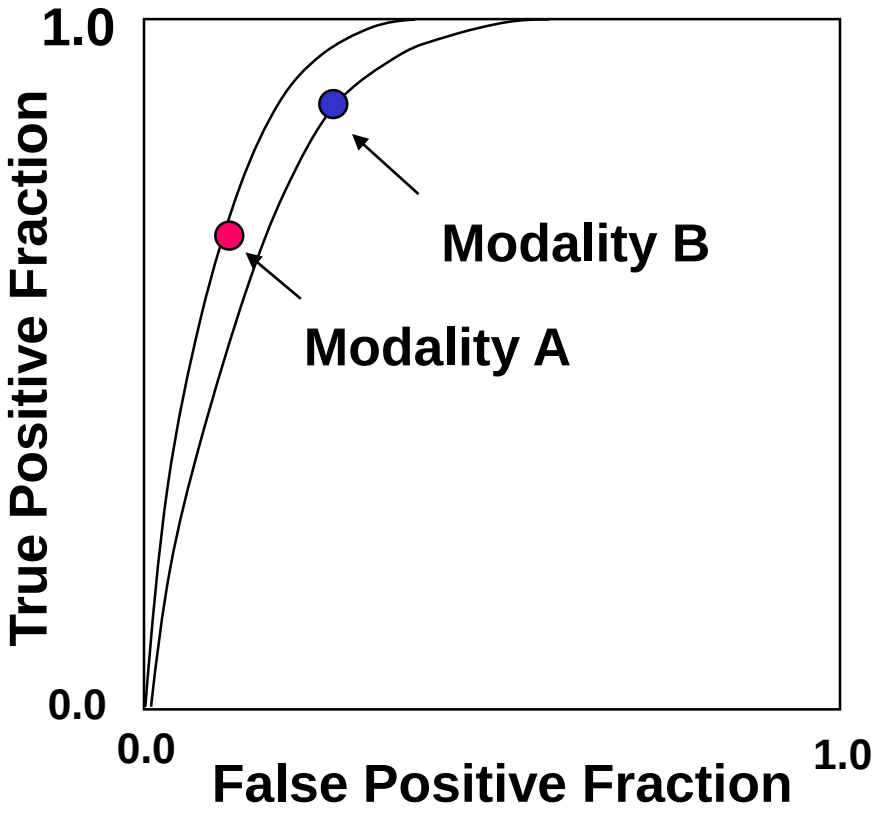
B is better than A

ROC
(another outcome)



B same as A

ROC
(another outcome)

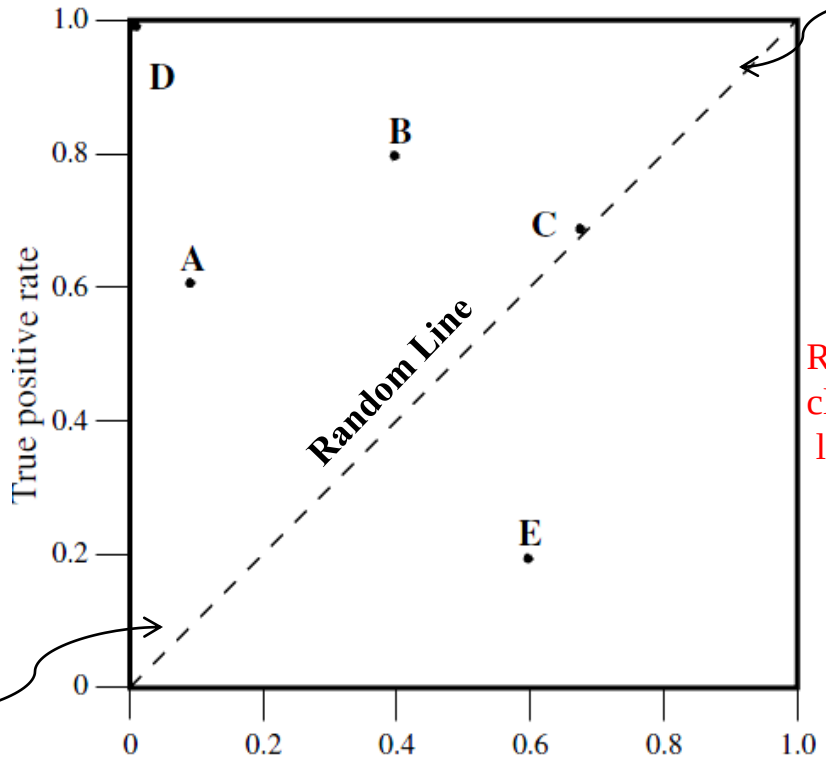


A is better than B

ROC Analysis

ROC graphs are two-dimensional graphs in which TPRate is plotted on the Y axis and FPRate is plotted on the X axis.
An ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives).

ROC graph with five classifiers labeled A through E.



Liberal: they make positive classifications with weak evidence. Classify nearly all positives correctly, but with high false positive rates.

ROC graph with five classifiers labeled A through E.

Conservative: they make positive classifications only with strong evidence so they make few false positive errors

ROC Analysis

Given an ROC graph in which a classifiers performance appears to be slightly better than random, it is natural to ask: ‘‘is this classifiers performance truly significant or is it only better than random by chance?’’

ROC Analysis

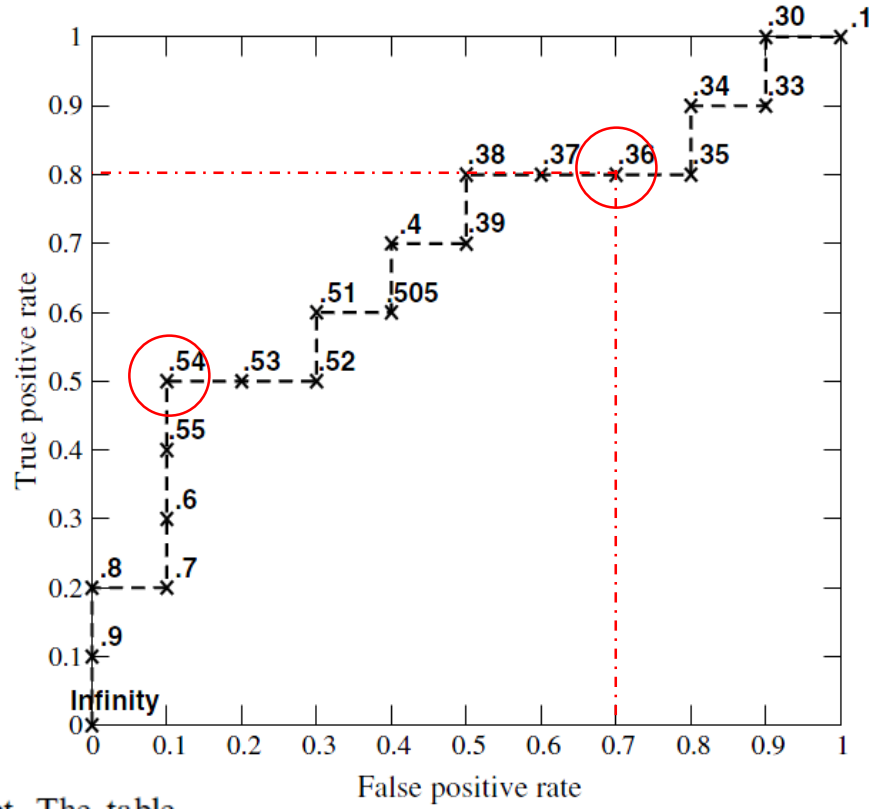
$$TPRate(0.36) = \frac{TP}{TP + FN} = \frac{8}{10} \qquad FPRate(0.36) = \frac{FP}{TN + FP} = \frac{7}{10}$$

$P = TP + FN = 10$; *Number of positives* $N = TN + FP = 10$; Number of negatives

Class p: Means instance is correctly classified with confidence (probability)=Score

Class n: Means instance is incorrectly classified with confidence (probability)=Score

	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. The table shows 20 data and the score assigned to each by a scoring classifier. The graph shows the corresponding ROC curve with each point labeled by the threshold that produces it.

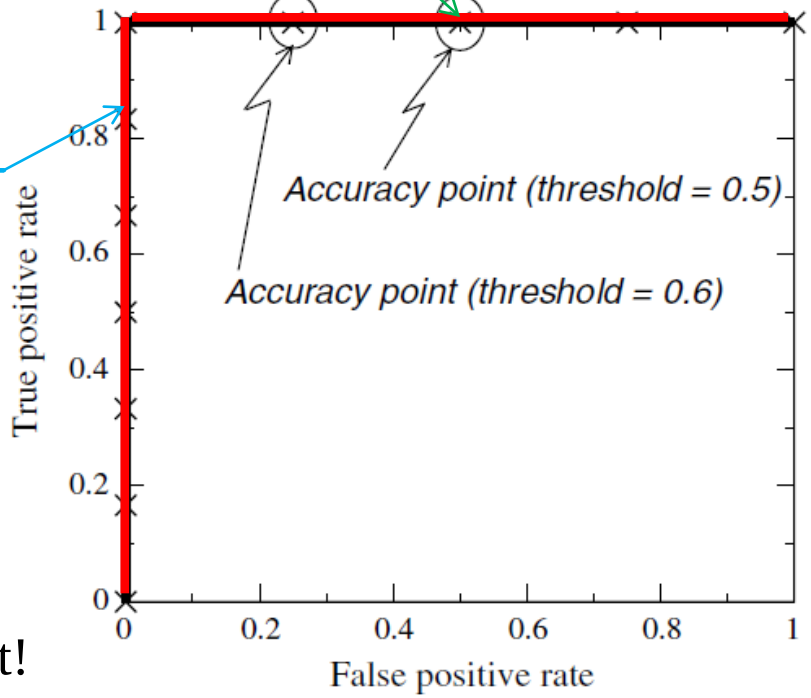
ROC Analysis – Relative vs. Absolute Scores

$TPRate(0.50961) = \frac{6}{6}$

$FPRate(0.50961) = \frac{2}{4}$

$TPRate(0.99964) = \frac{5}{6}$

$FPRate(0.99964) = \frac{0}{4}$



Inst no.	Class True	Class Hyp	Score
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

Scores and classifications of 10 instances, and the resulting ROC curve.

Is it ideal? Is not!