MIA – Medical Image Analysis

Evaluating your algorithm

Master VIBOT – University of Girona

2016-2017

CAD: Computer-Aided Diagnosis

- Evaluation / Classification

  - CAD systems seek to highlight suspicious structures.
  - The aim is to properly **classify** each object instance to a particular class.
  - Thus, perform a process that **assigns a label** to an object according to some representation of the object's properties.

CAD: Computer-Aided Diagnosis

- Evaluation / Classification (some definitions)
  - Classifier: device or algorithm that inputs an object representation and outputs a class label.
  - Reject class: generic class for objects that cannot be placed in any of the designated known classes.
  - Classification error = misclassification
  - Empirical error rate = % misclassified
  - Empirical reject rate = % rejected

CAD: Computer-Aided Diagnosis

- Evaluation / Classification (some definitions)
  - Training set
    - Sample data for which truth is known (true label).
    - Used to develop the classifier.
  - Testing set (independent test data)
    - Sample data
    - Not part of training set
    - Sampled from real population (ideally)

CAD: Computer-Aided Diagnosis

- Evaluation / Classification: two-class problems
  - Good / Bad
  - Present / Absent
  - Disease / Not disease
  - Metal / Not metal
  - …

truth	classifier	evaluation
+	+	**true positive**
-	+	**false positive**
-	-	**true negative**
+	-	**false negative**

Two types of error:

False positive (false alarm), **FP**:
Alarm sounds, but is not carrying metal

False negative (miss), **FN**:
Alarm doesn't sound, but is carrying metal!

CAD: Computer-Aided Diagnosis

- Classifier performance: two-class problems
  - Let's consider classification problems using only two classes:
    - Each instance ***I*** is mapped to one element of the set {***p,n***} of positive and negative class labels.
    - A classification model (or classifier) is a mapping from instances to predicted classes.
    - We use the labels {***Y,N***} for the class predictions produced by a model.
    - Given a classifier and an instance, there are four possible outcomes:
      - True positive: the instance is positive and it is classified as positive
      - False negative: the instance is positive and it is classified as negative
      - True negative: the instance is negative and it is classified as negative
      - False positive: the instance is negative and it is classified as positive

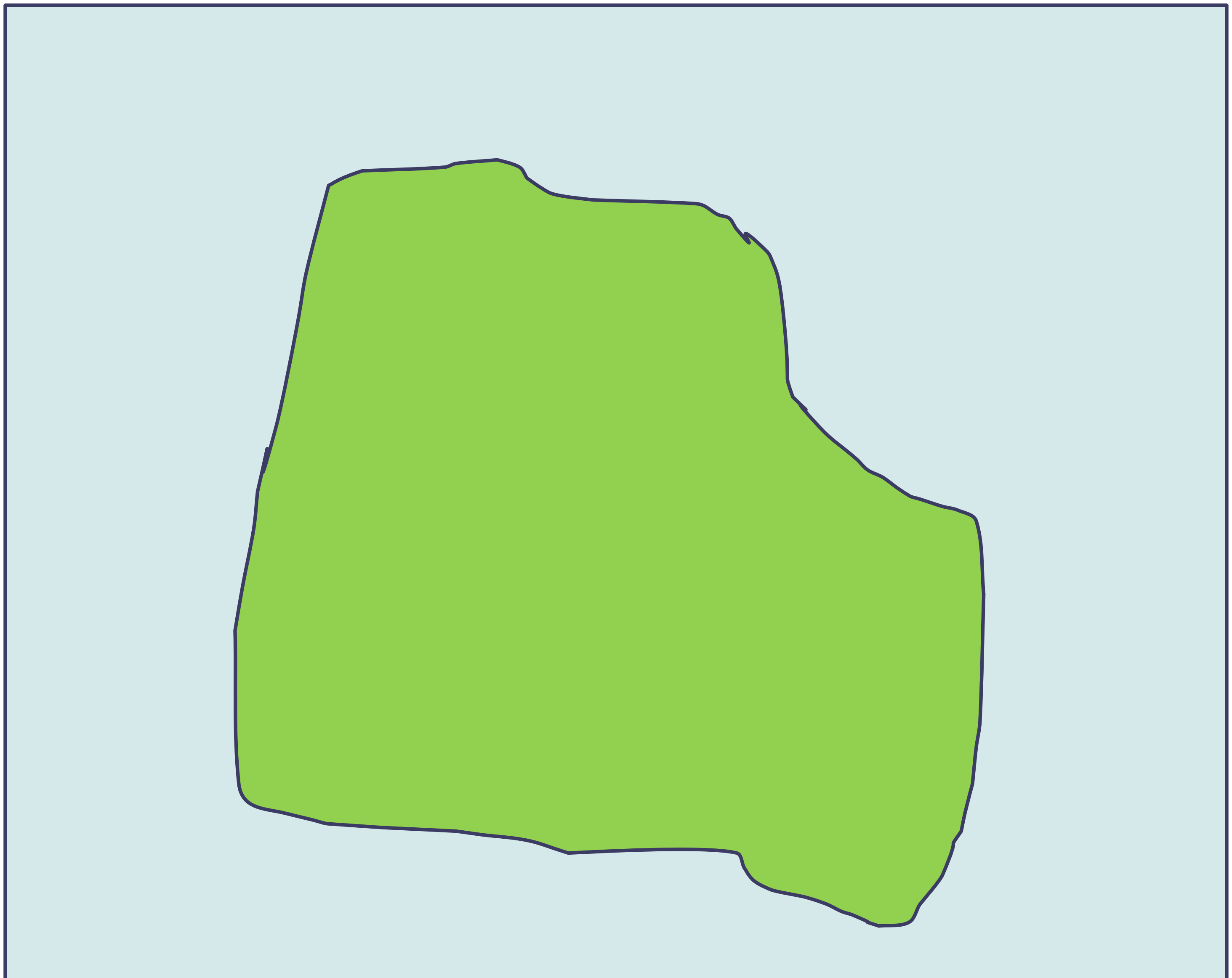
CAD: Computer-Aided Diagnosis

- Classifier performance: two-class problems
  - Given a classifier and a set of instances (the test set), a two-by-two **confusion matrix** can be constructed representing the dispositions of the instances.

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives

  - The numbers along the major diagonal represent the correct decisions made.
  - The numbers outside the major diagonal represent the errors (the confusion) between the various classes.

CAD: Computer-Aided Diagnosis



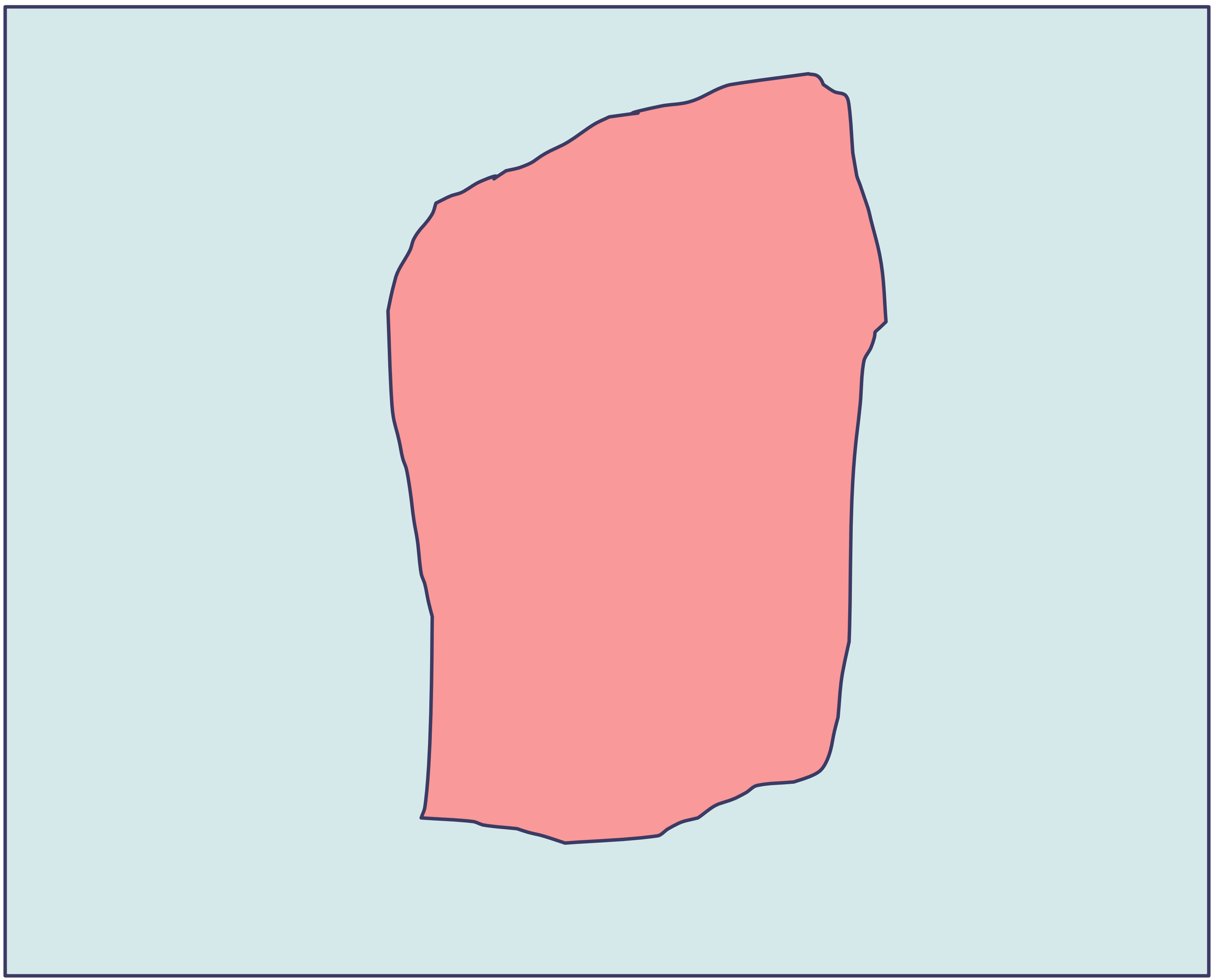
True Class (truth)

Positive

Negative

Hypothesized class		True class: p	True class: n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives

CAD: Computer-Aided Diagnosis

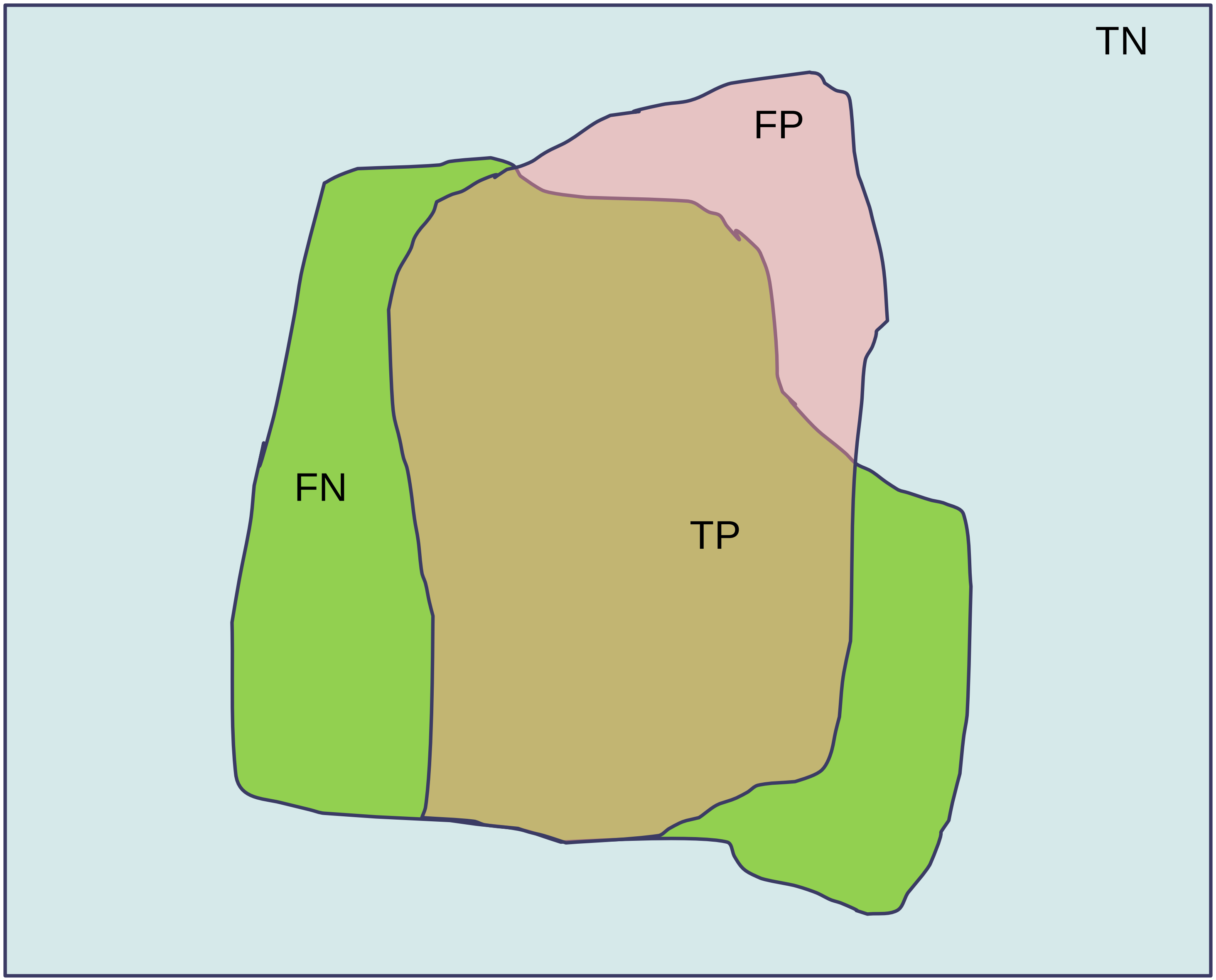


Hypothesized Class (classifier)

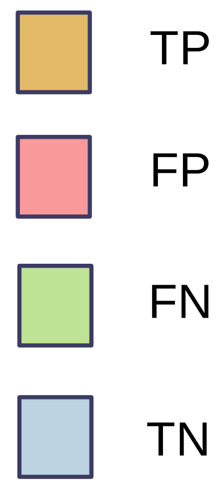
- Yes prediction
- Not prediction

Hypothesized class		True class: p	True class: n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives

CAD: Computer-Aided Diagnosis



Comparing True to Hypothesized Class

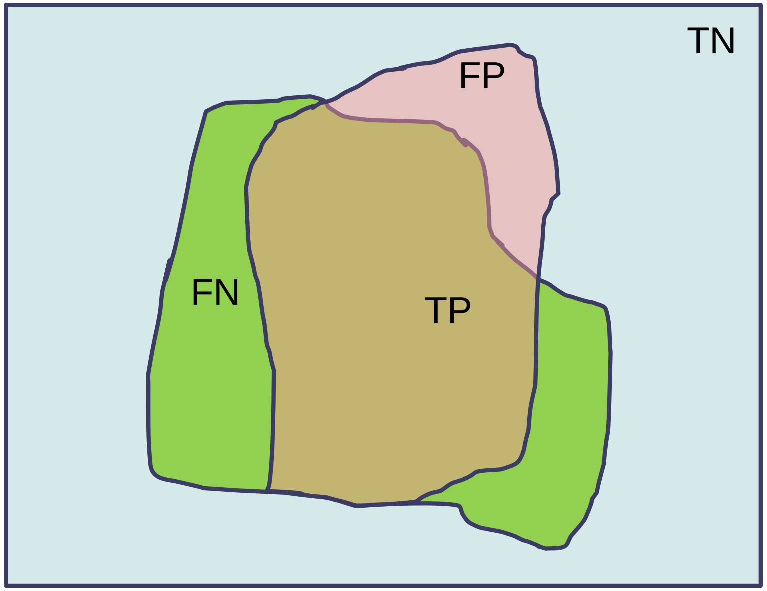


		True class p	True class n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives

CAD: Computer-Aided Diagnosis

- Classifier performance: two-class problems
  - Given a classifier and a set of instances (the test set), a two-by-two **confusion matrix** can be constructed representing the dispositions of the instances.

		True class	
		p	n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives



  - The numbers along the major diagonal represent the correct decisions made.
  - The numbers outside the major diagonal represent the errors (the confusion) between the various classes.

CAD: Computer-Aided Diagnosis

- Classifier performance: evaluating errors on two-class problems
  - Diagnostic accuracy: agreement between diagnoses and “truth”
  - Different costs are associated with different errors
  - Ex. Mammography
    - Let’s suppose 99% of population holds for N.A.D. (no apparent disease)
      - The algorithm works 99% of the time.
      - However, the cost of 1% misclassified cases is great (death).
    - On the contrary, let’s suppose disease is always present:
      - The classification algorithm is wrong 99% of the time!
      - It is need to follow up with biopsy.
      - Psychological stress will probably be inflicted on many patients.

CAD: Computer-Aided Diagnosis

- Classifier performance: accuracy
  - For a single t (decision threshold) setting, we can calculate (test our method for) TP, TN, FP, and FN.
  - TP + TN + FP + FN = # of normal and abnormal cases in the study population.
  - **Accuracy**: (TP + TN) / (TP + TN + FP + FN)
    - This is greatly influenced by target (or disease) prevalence.
  - Ex. Mammography
    - The disease occurs > 1%
    - So, let a trained mo[...]s say there is no disease (even without looki[...]igital mammogram!)
    - Our mo[...]ccuracy is (0+99)/0+99+0+1 = 99%

We need a different measure!

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - "The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research."

http://en.wikipedia.org/wiki/Receiver_operating_characteristic

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - “Following the attack on Pearl Harbor in 1941, the United States Army began new research to increase the prediction of correctly detected Japanese aircraft from their radar signals.”

http://en.wikipedia.org/wiki/Receiver_operating_characteristic




CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - "In the 1950s, ROC curves were employed in psychophysics to assess human (and occasionally non-human animal) detection of weak signals. In medicine, ROC analysis has been extensively used in the evaluation of diagnostics tests."
  - "In radiology, ROC analysis is a common technique to evaluate new radiology techniques".
  - "ROC curves also proved useful for the evaluation of machine learning techniques, in comparing and evaluating different classification algorithms".

http://en.wikipedia.org/wiki/Receiver_operating_characteristic

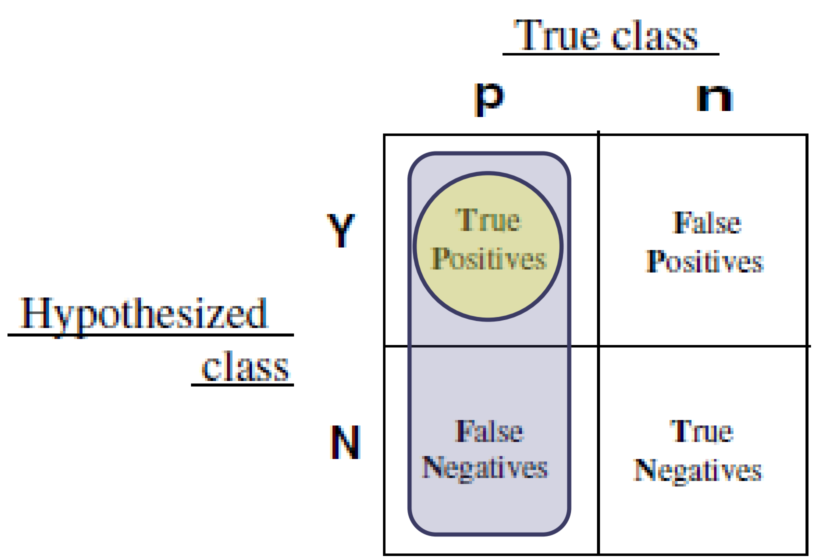
CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - **Sensitivity**: the number of abnormal cases correctly identified divided by the total number of abnormal cases -> relates to the test's ability to **identify positive results.**
  - **Specificity**: the number of normal cases correctly identified as negative divided by the total number of normal cases -> relates to the test's ability to **identify negative results.**
  - Use of ROC (Receiver Operating Characteristics) graph.
  - Distance measuring in clustering.

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - **Sensitivity**: the number of abnormal cases correctly identified divided by the total number of abnormal cases (relates to the test's ability to **identify positive results**).

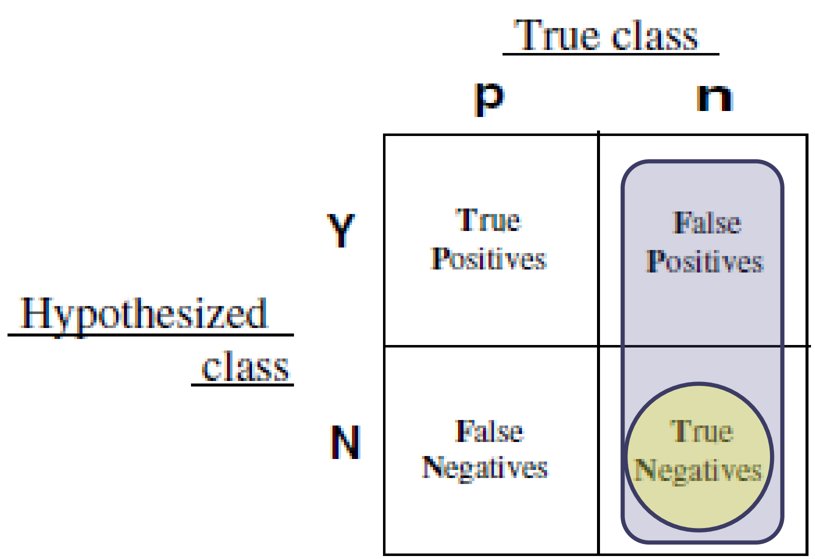
$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$



CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - **Specificity**: the number of normal cases correctly identified as negative divided by the total number of normal cases (relates to the test's ability to **identify negative results**).

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$



CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Common evaluation metrics from the confusion matrix:

$$tp\ rate \approx \frac{\text{Positives correctly classified}}{\text{Total positives}}$$

$$tp\ rate = \frac{TP}{P} = \frac{TP}{TP + FN} = recall$$

$$fp\ rate \approx \frac{\text{Negatives incorrectly classified}}{\text{Total negatives}}$$

$$fp\ rate = \frac{FP}{N} = \frac{FP}{FP + TN}$$

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

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

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

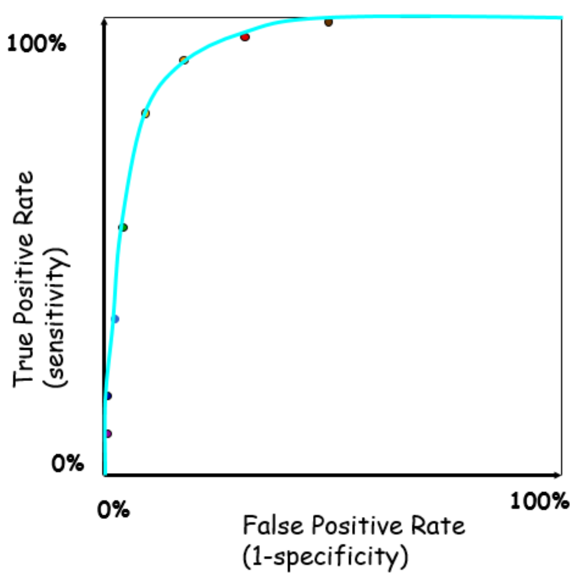
$$precision = \frac{TP}{Y} = \frac{TP}{TP + FP}$$

Hypothesized class		True class: p	True class: n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives

CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - A ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance.
  - **ROC** graphs have long been used in signal detection theory to depict the **tradeoff between hit rates and false alarm rates of classifiers**.
  - When **evaluating binary decision problems**, ROC curves show how the number of correctly classified positive examples varies with the number of incorrectly negative examples (**TPR vs FPR**).

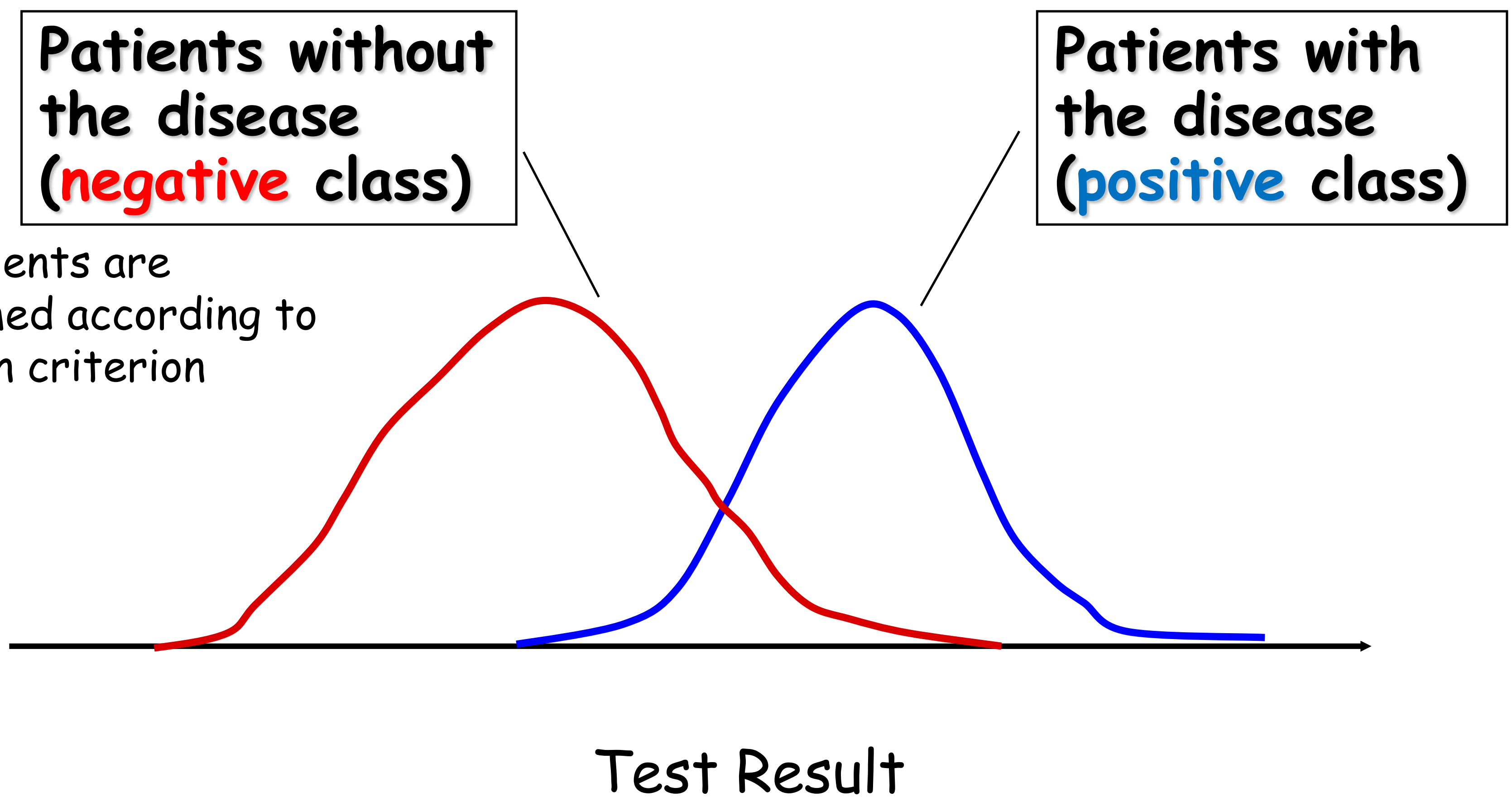
ROC curve



CAD: Computer-Aided Diagnosis

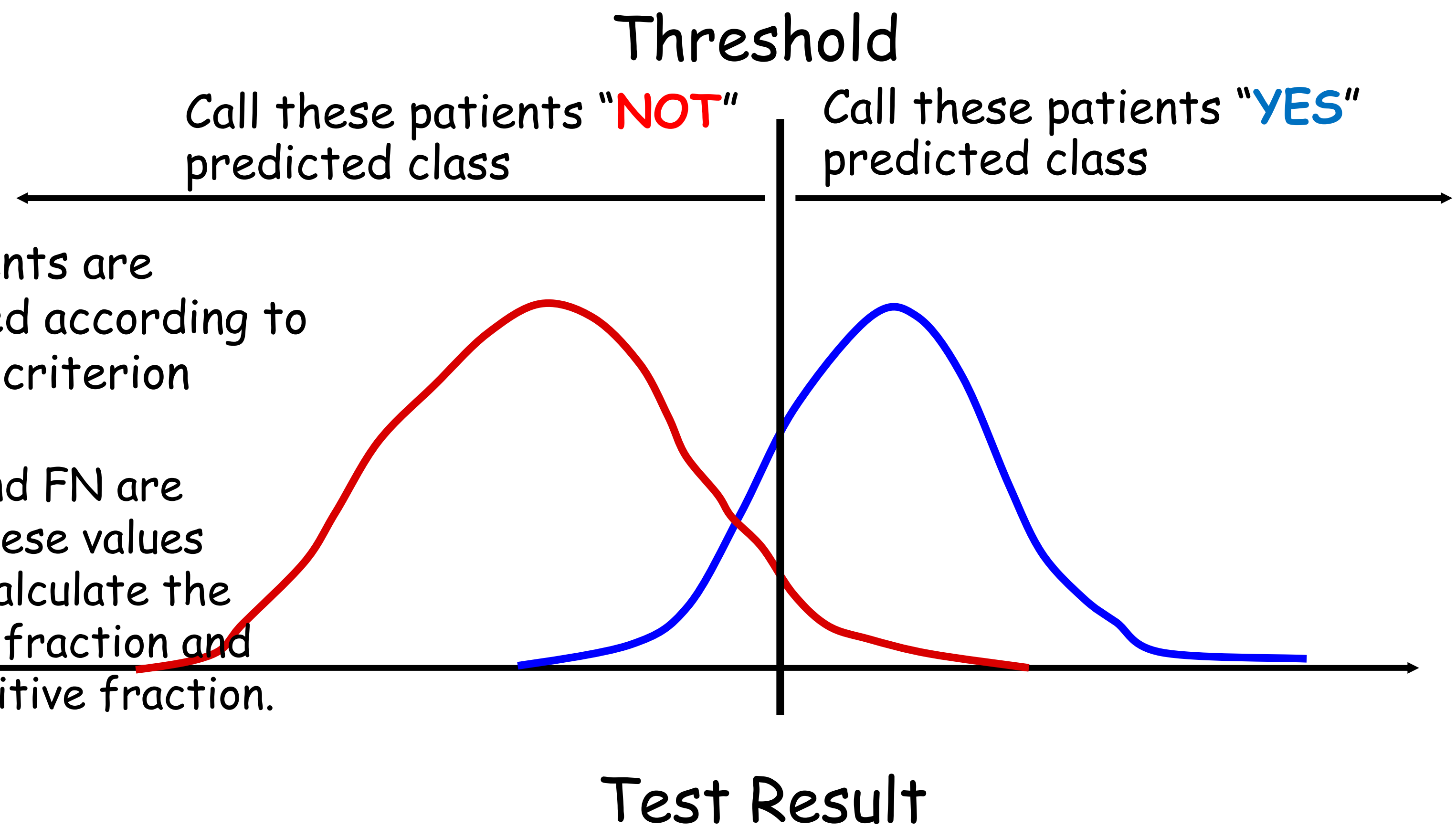
- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example

n and p patients are histogrammed according to the decision criterion



CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



n and p patients are histogrammed according to the decision criterion

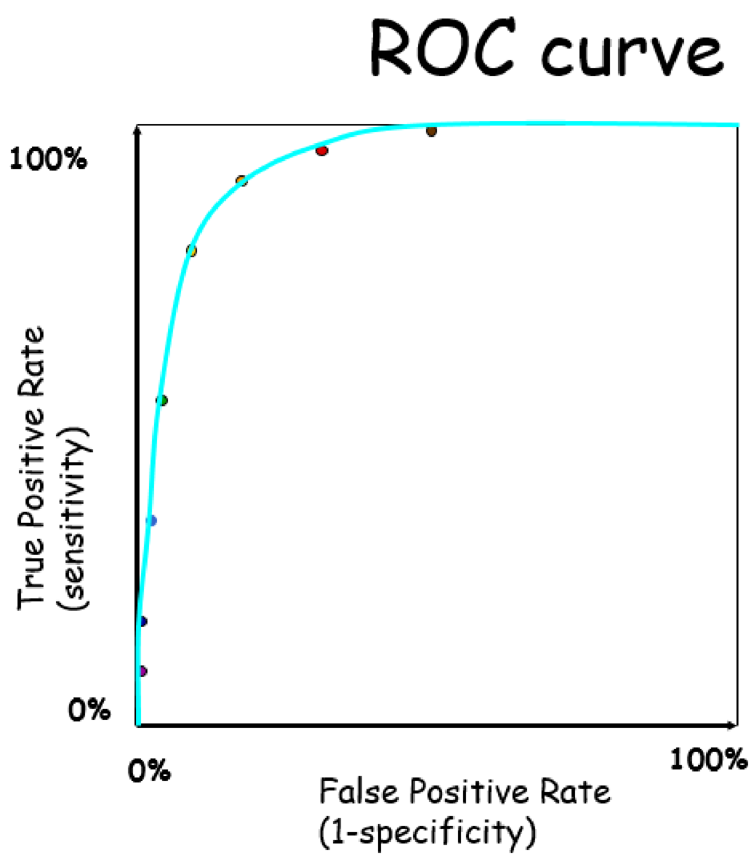
TP, TN, FP and FN are computed. These values are used to calculate the true-positive fraction and the false-positive fraction.

CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - A ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance.

n and p patients are histogrammed according to the decision criterion

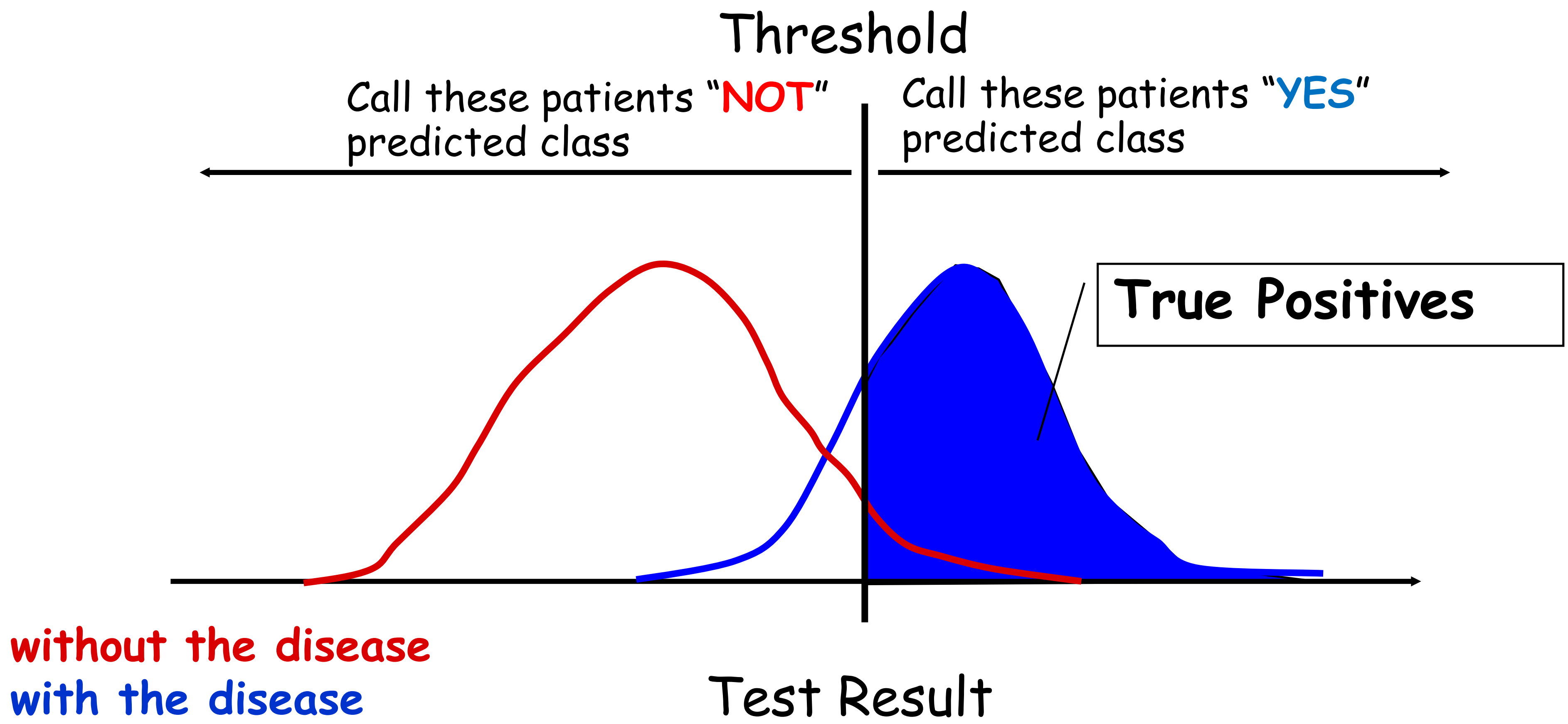
TP, TN, FP and FN are computed. These values are used to calculate the true-positive fraction and the false-positive fraction.



The decision threshold at each point gives rise to one point on the ROC curve. Shifting the decision threshold towards the left will increase the true-positive fraction but will also increase the false-positive fraction.

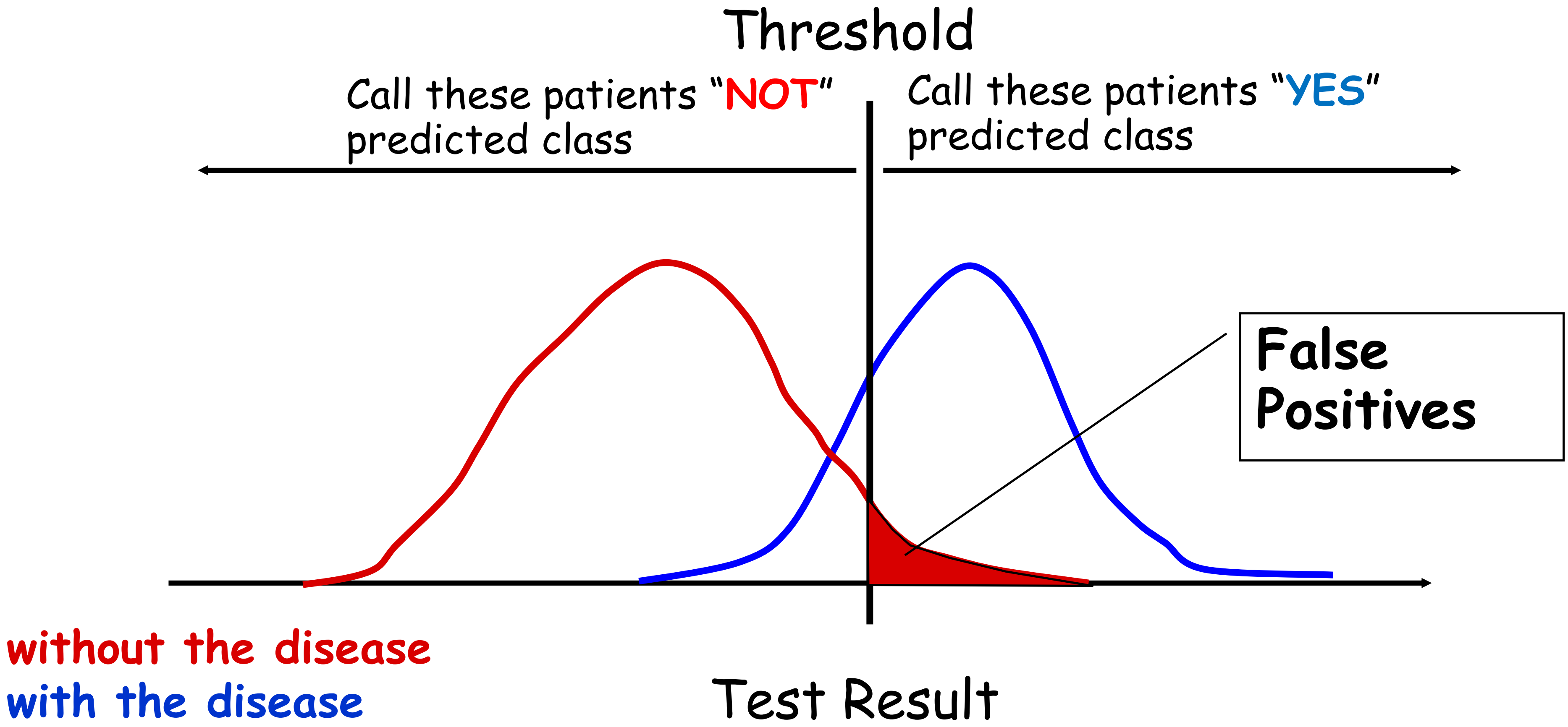
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



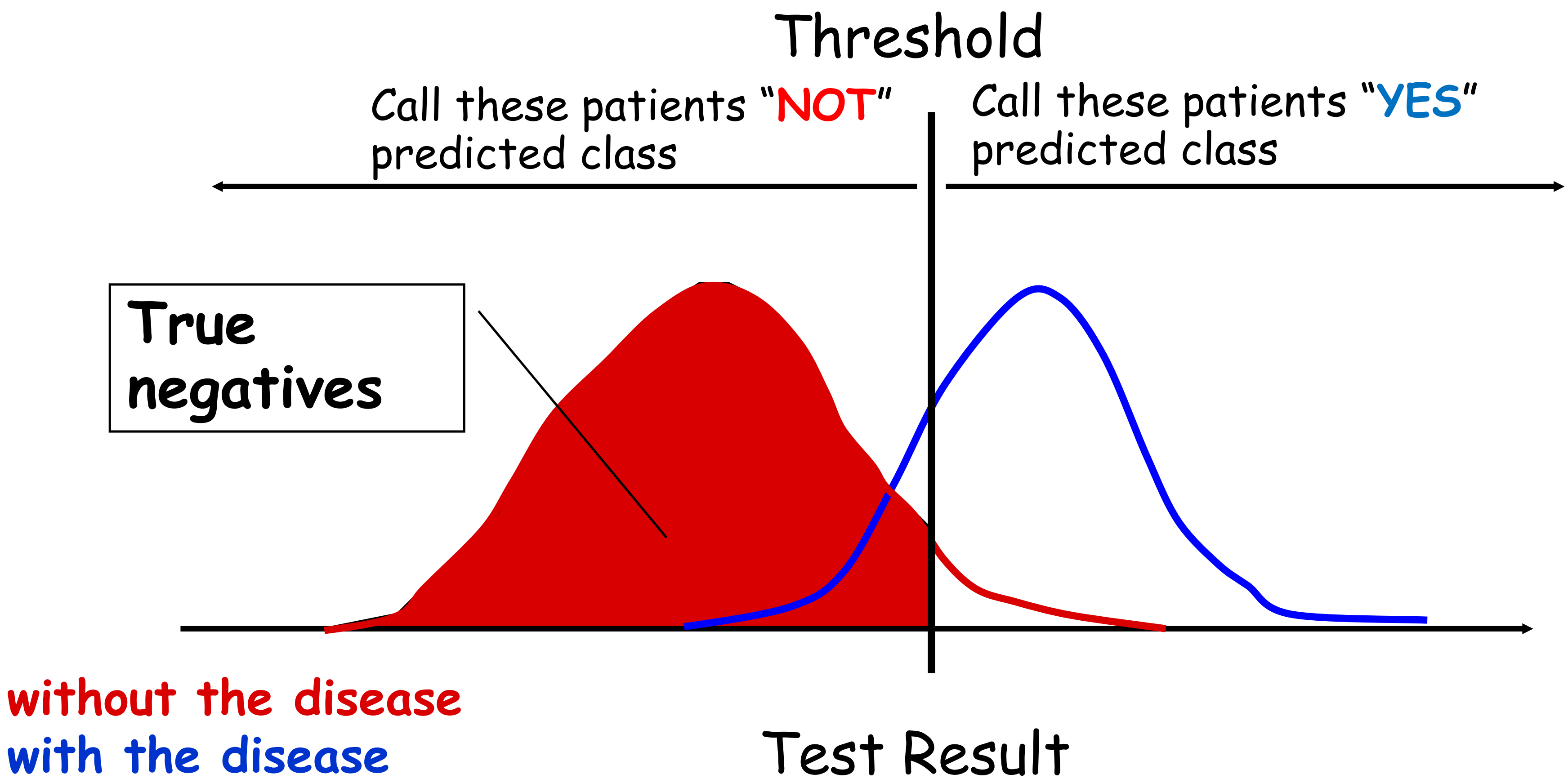
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



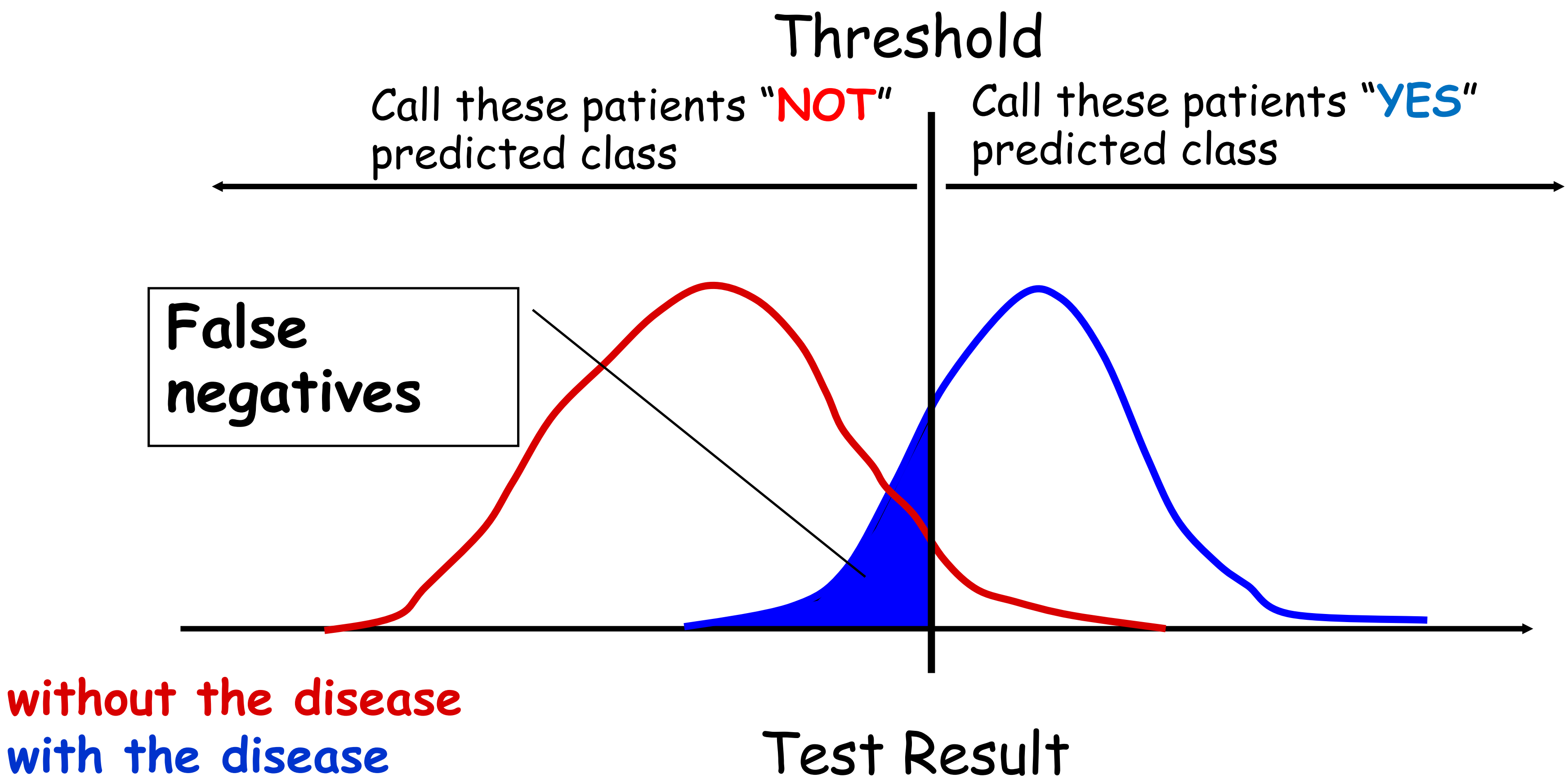
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



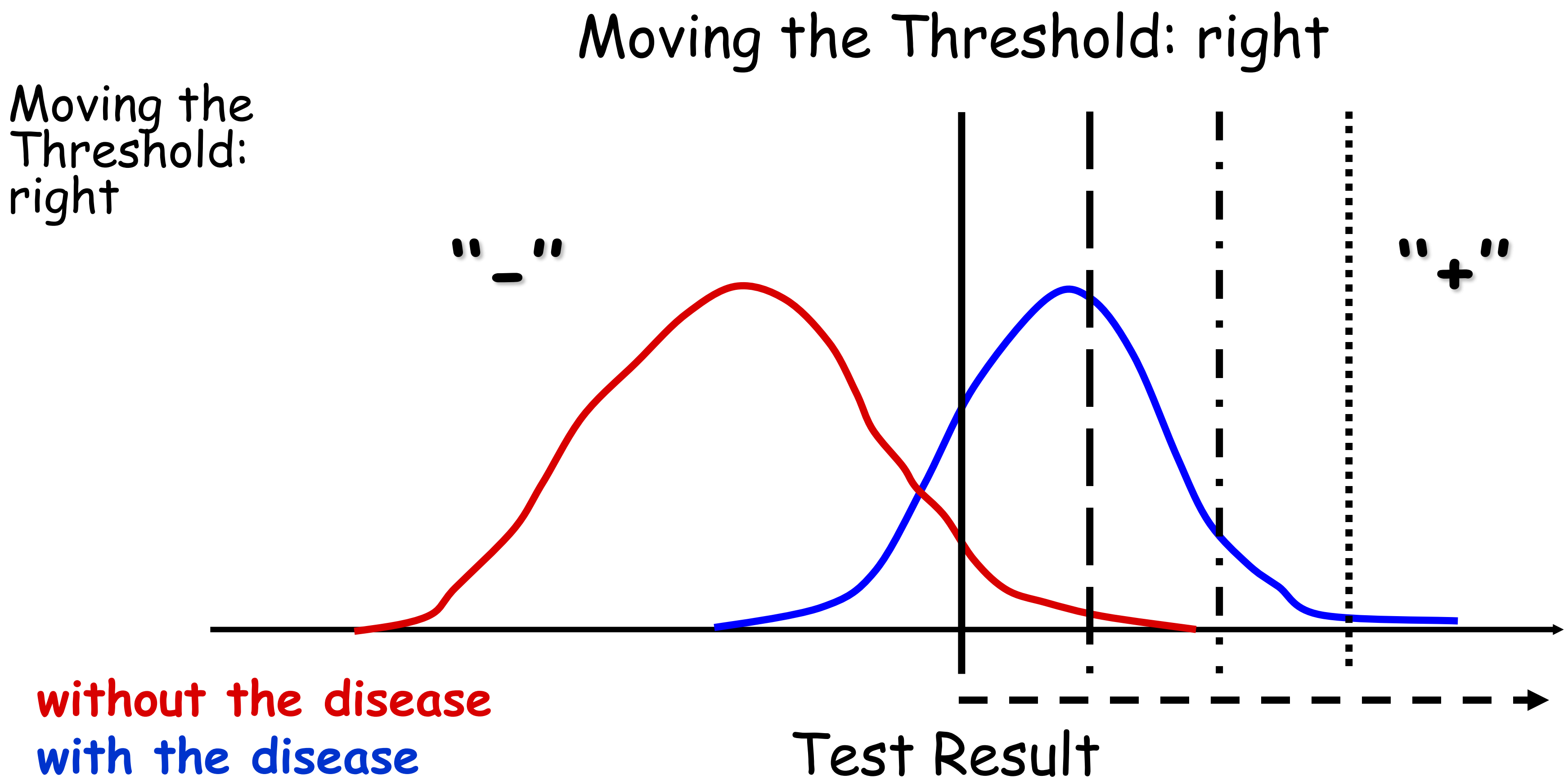
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



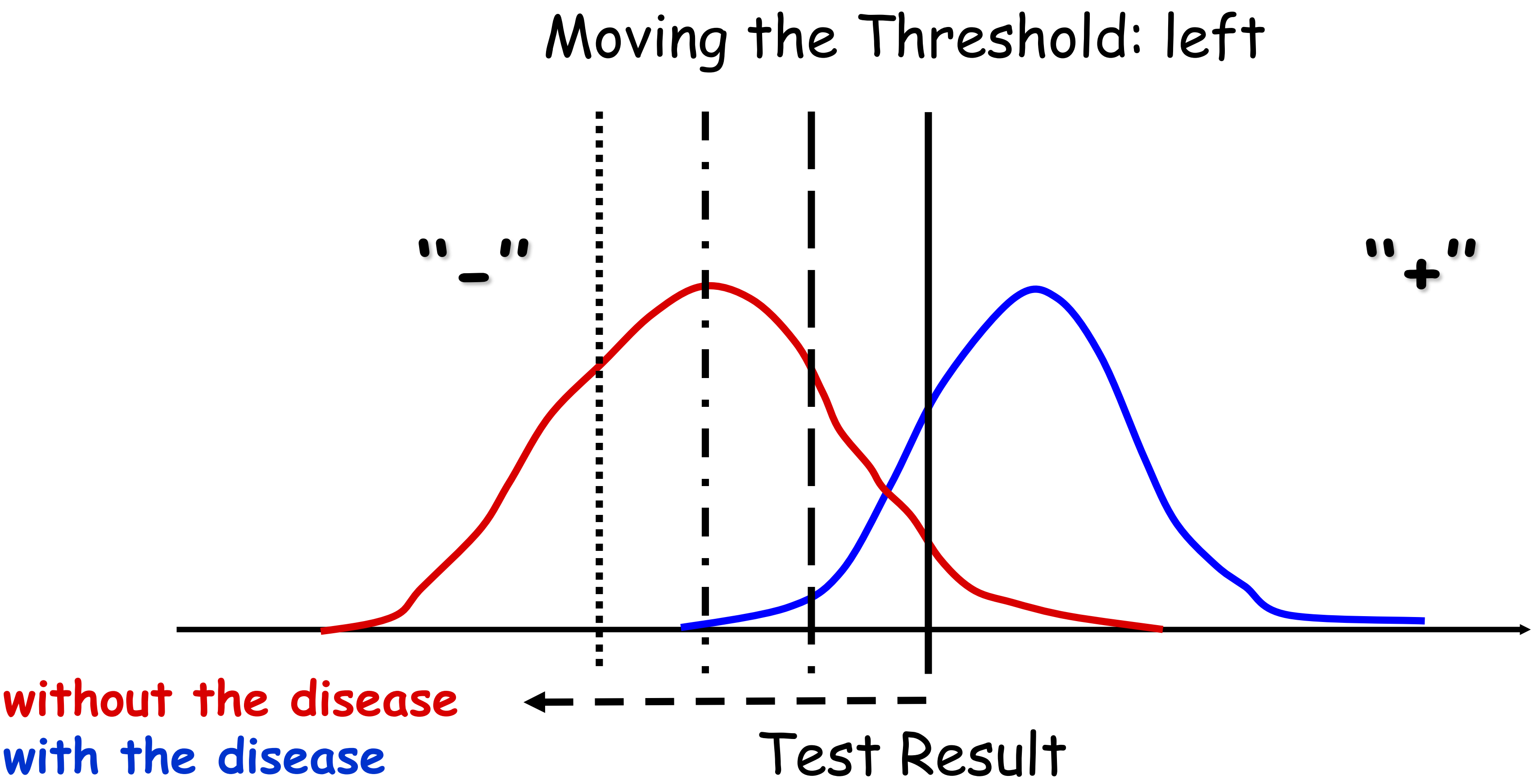
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



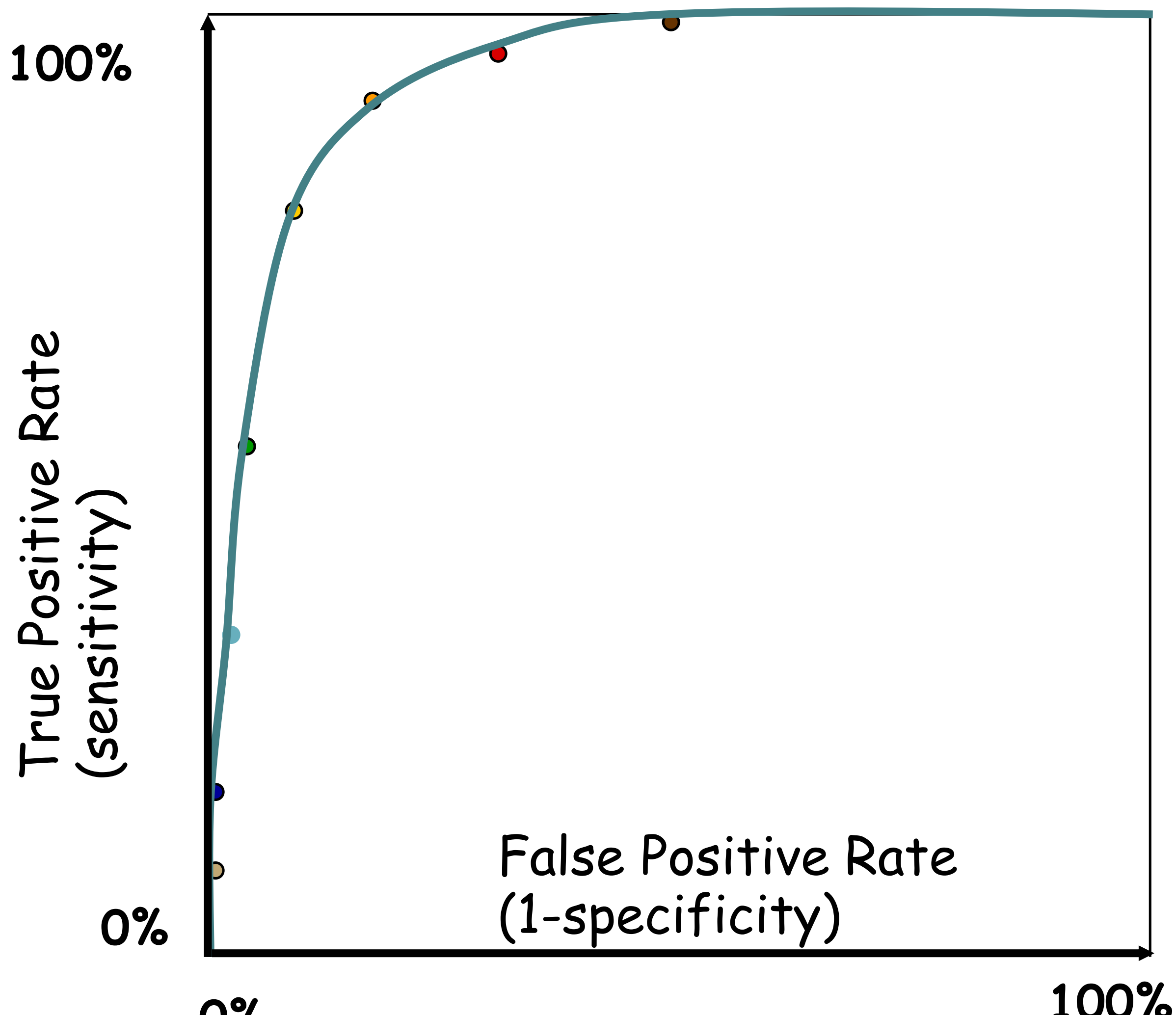
CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC Analysis
  - The basics of signal detection theory: example



CAD: Computer-Aided Diagnosis

- Evaluation / Classification: ROC curve



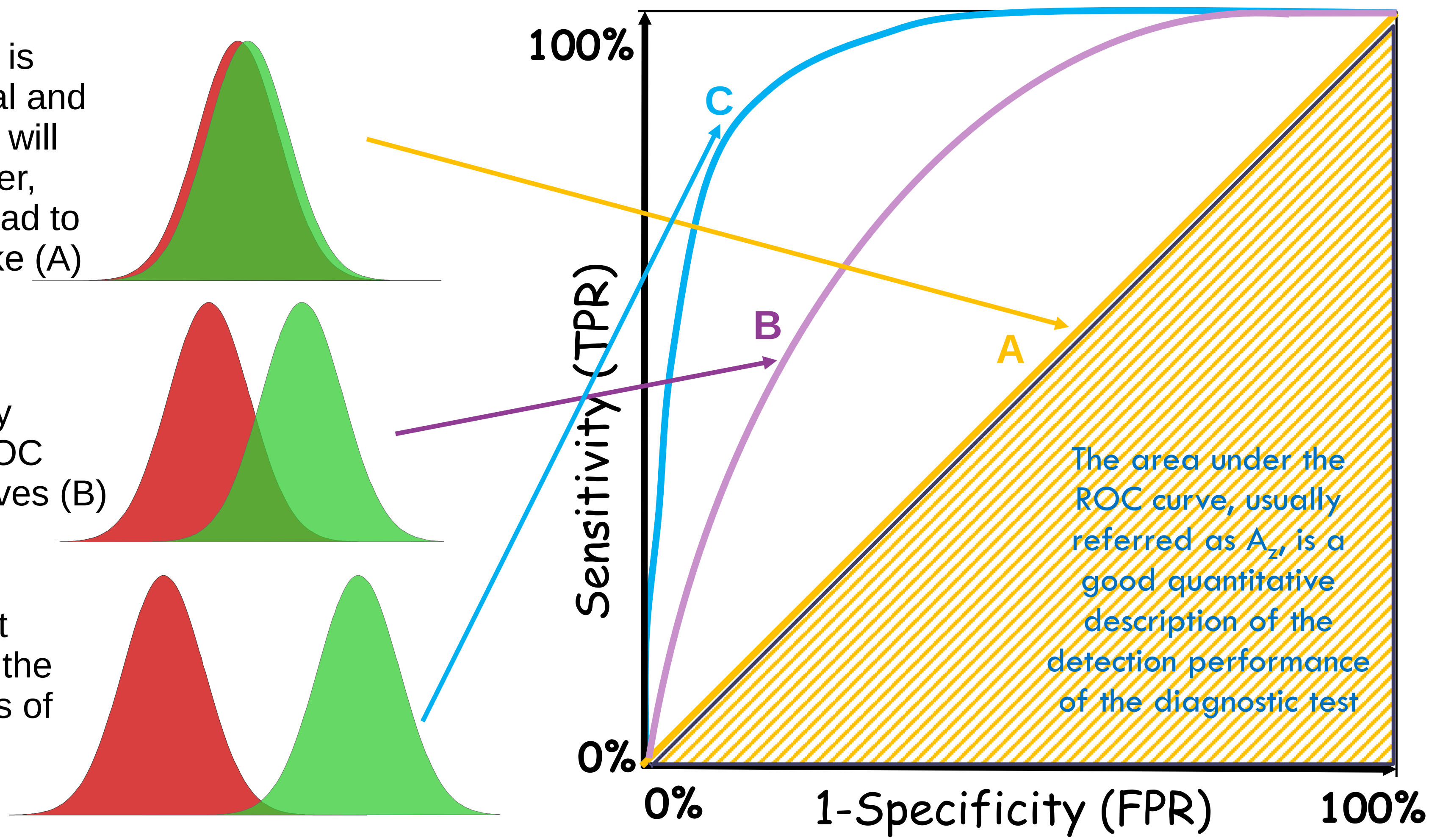
CAD: Computer-Aided Diagnosis

- ROC curve: evaluating classifiers

If the separability between classes is close to 0, normal and abnormal curves will overlay each other, and this would lead to an ROC curve like (A)

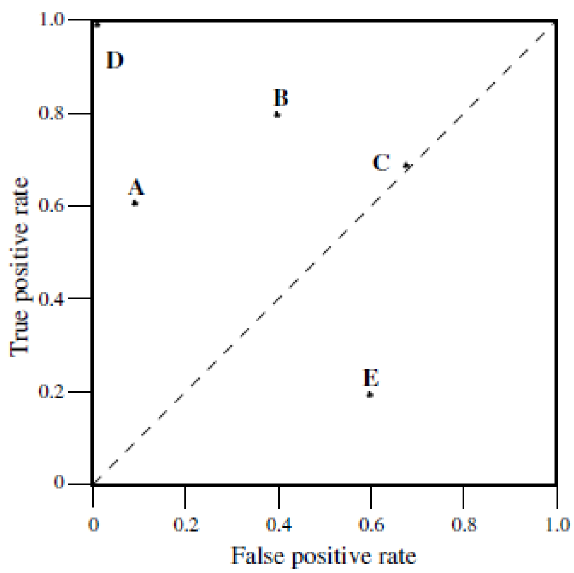
As the separability increases, the ROC curve also improves (B)

An almost-perfect ROC curve rides the left and top edges of the ROC plot (C)



CAD: Computer-Aided Diagnosis

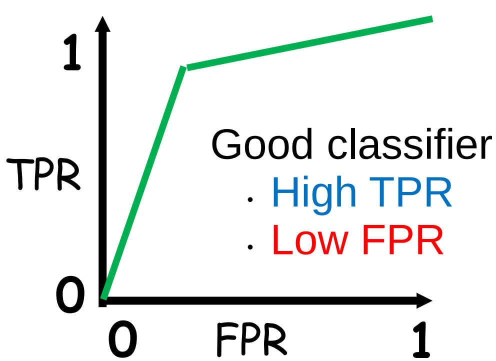
- Classifier performance: ROC Analysis
  - A discrete classifier is one that outputs only a class label.
  - Each discrete classifier produces an (*fp rate*, *tp rate*) pair corresponding to a single point in ROC space.

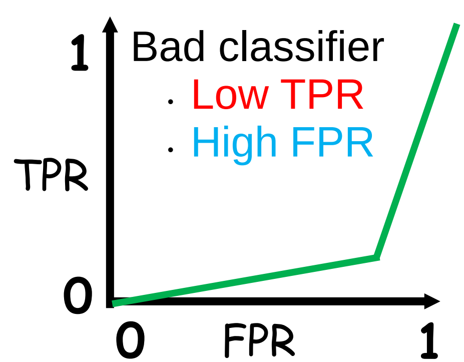


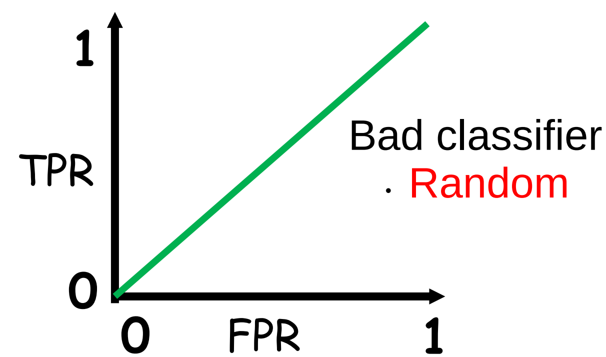
A basic ROC graph showing five discrete classifiers.

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Several important points in ROC space:
    - (0,0): Never issuing a positive classification. No false positive errors, but also gains no true positives.
    - (1,1): Unconditionally issuing positive classifications (the opposite of previous strategy).
    - (0,1): Perfect classification
    - Informally, one point in ROC space is better than another if it is to the northwest (*tp rate* is higher, *fp rate* is lower, or both) of the first.







CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Random performance
    - The diagonal line $y = x$ represents the strategy of randomly guessing a class.
    - For example, if a classifier randomly guesses the positive class half the time, it can be expected to get half the positives and half the negatives correct; this yields the point (0.5,0.5) in ROC space.
    - If it guesses the positive class 90% of the time, it can be expected to get 90% of the positives correct but its false positive rate will increase to 90% as well, yielding (0.9,0.9) in ROC space.
    - Thus, a random classifier will produce a ROC point that slides back and forth on the diagonal based on the frequency with which it guesses the positive class.

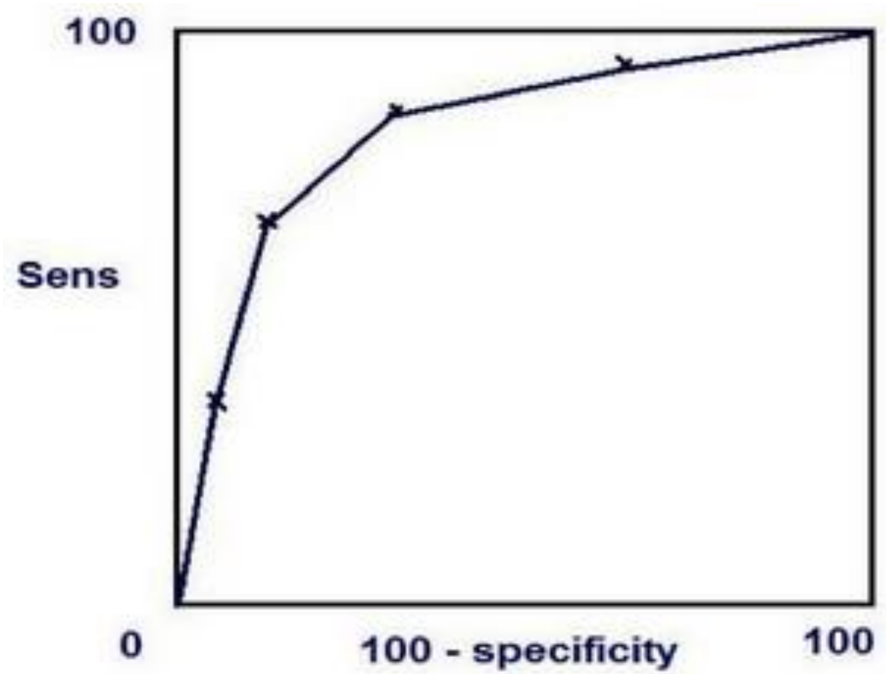
CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Curves in ROC space
    - A discrete classifier produces only a single point in ROC space.
    - Some classifiers (e.g. a Naive Bayes classifier or a Neural Network) yield an instance probability or score, a numeric value that represents the degree to which an instance is a member of a class.
    - Such ranking or scoring classifier can be used with a threshold to produce a discrete (binary) classifier: if the classifier output is above the threshold, the classifier produces a ***Y***, else a ***N***.
    - Each threshold value produces a different point in ROC space. Conceptually, we may imagine varying a threshold from $-\infty$ to $+\infty$ and tracing a curve through ROC space.

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis (example)
  - Curves in ROC space: detecting German planes in WWII

Radar detector setting	Percent of German planes detected (**sensitivity**)	Percent of geese flocks correctly identified (**specificity**)	Percent of geese flocks incorrectly identified (**1-specificity**)
Off	0	100	0
Setting 1	35	93	7
Setting 2	60	85	15
Setting 3	85	70	30
Setting 4	92	30	70
Full	100	0	100



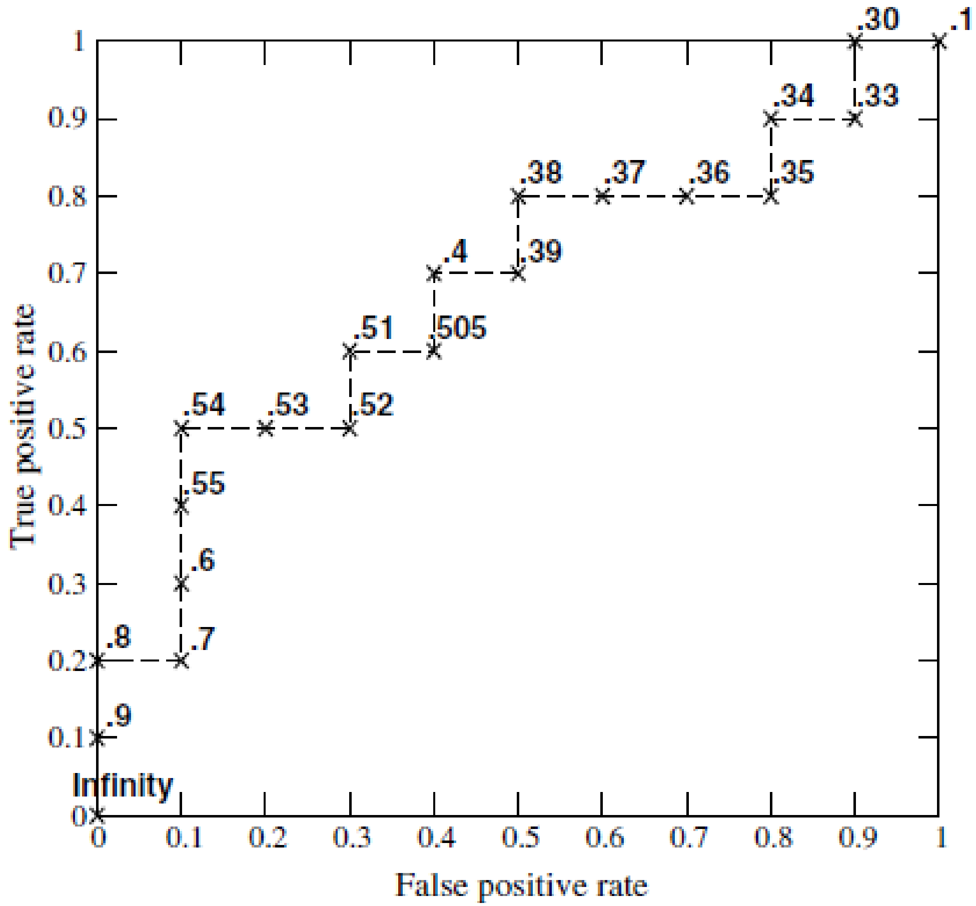
As we turn the sensitivity of the receiver up, it detects more and more German planes but also mislabels more flocks of geese as German planes. So, as sensitivity goes up, specificity goes down.

Instead of radar, German planes and geese flocks, in CAD systems let's talk about diagnosis tests (mammograms), disease-findings (malignant masses) and non-disease findings (benign masses).

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Curves in ROC space: example on a test set of 20 instances

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



CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Curves in ROC space: example on a test set of 20 instances
    - TP and FP both start at zero.
    - TP rate axis (Y) will be divided into as many bins as positive instances the dataset contains.
    - FP rate axis (X) will be divided into as many bins as negative instances the dataset contains.
    - For each positive instance we increment TP and for every negative instance we increment FP.

CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Curves in ROC space: comparing GT with algorithm results

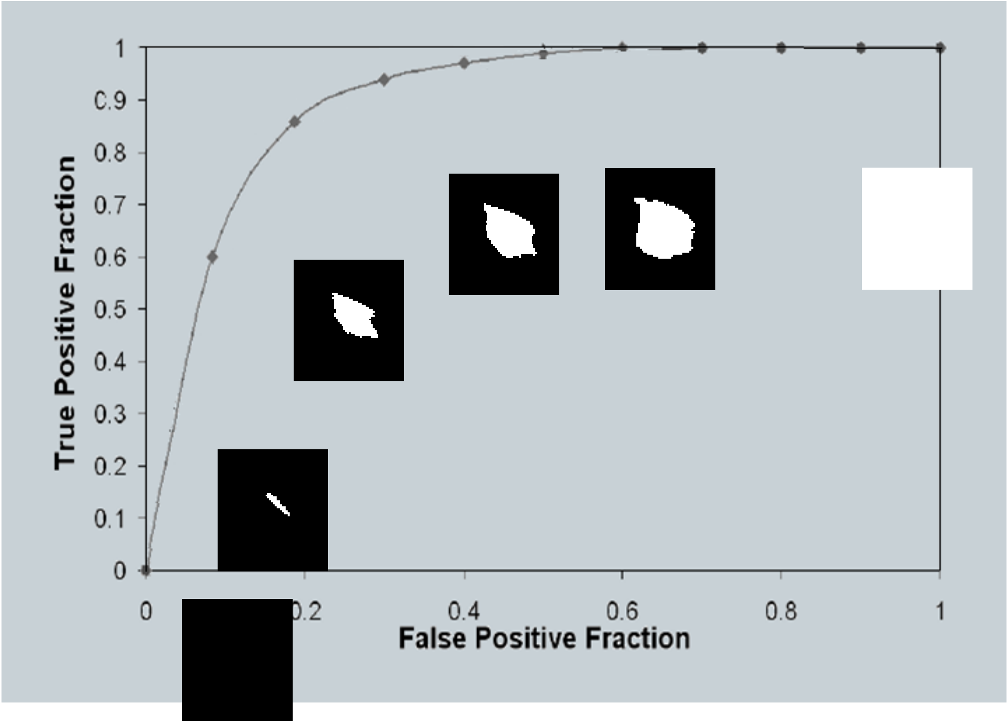
Algorithm \ doctor	Mass	No Mass
Mass	TP	FP
No Mass	FN	TN
	↓ Sensitivity	↓ Specificity

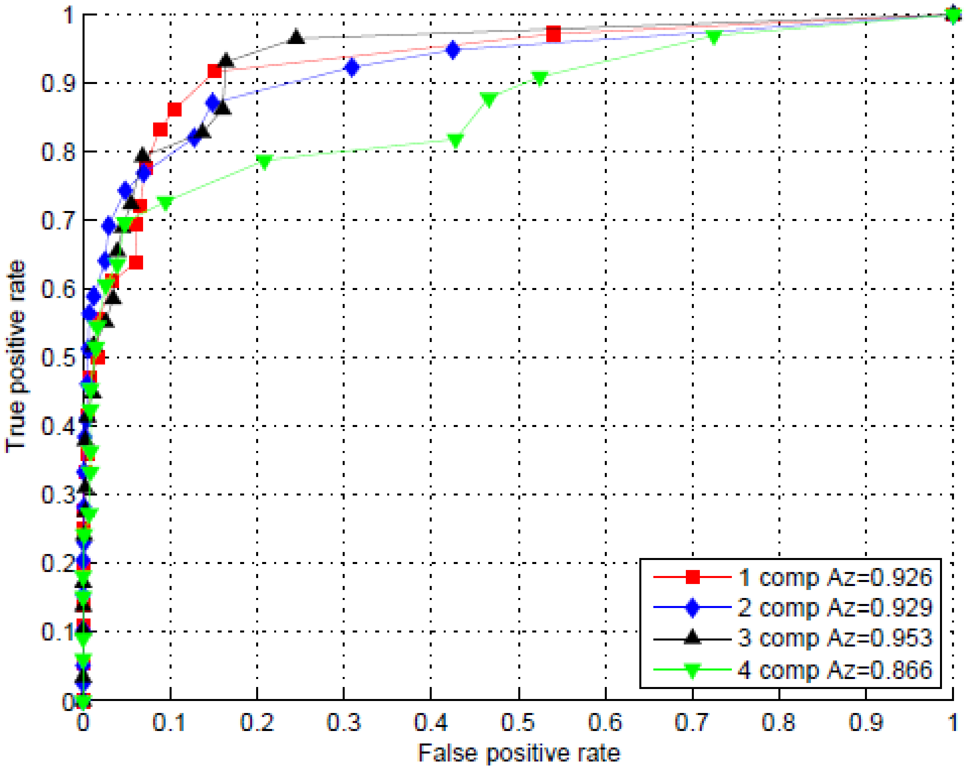
$$\text{sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}.$$

$$\text{specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

CAD: Computer-Aided Diagnosis

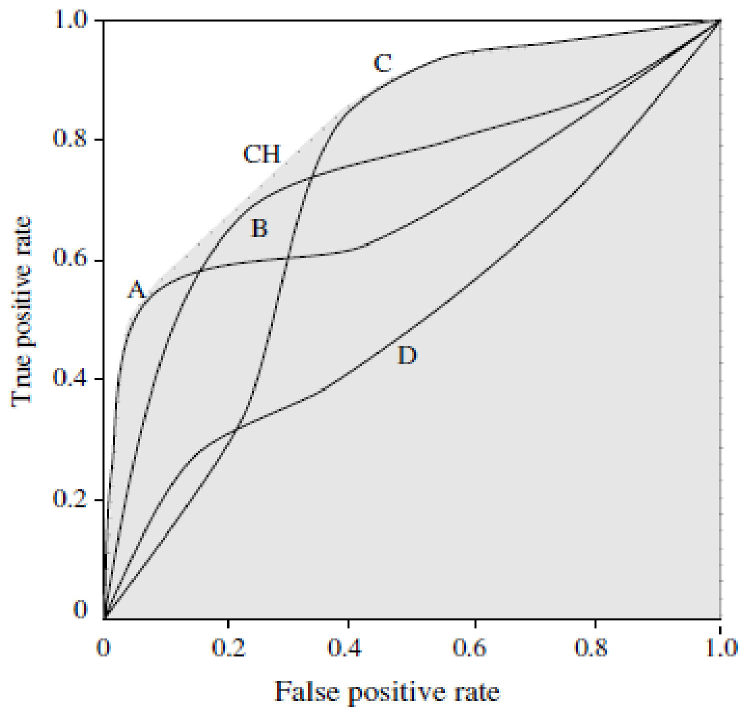
- Classifier performance: ROC Analysis
  - Curves in ROC space: comparing GT with algorithm results





CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Curves in ROC space: comparing different algorithms (diagnoses): which one is better?
  - Sensitivity and specificity don't always make it clear which of two diagnostic tests is better
  - Area under the curve (**AUC**)

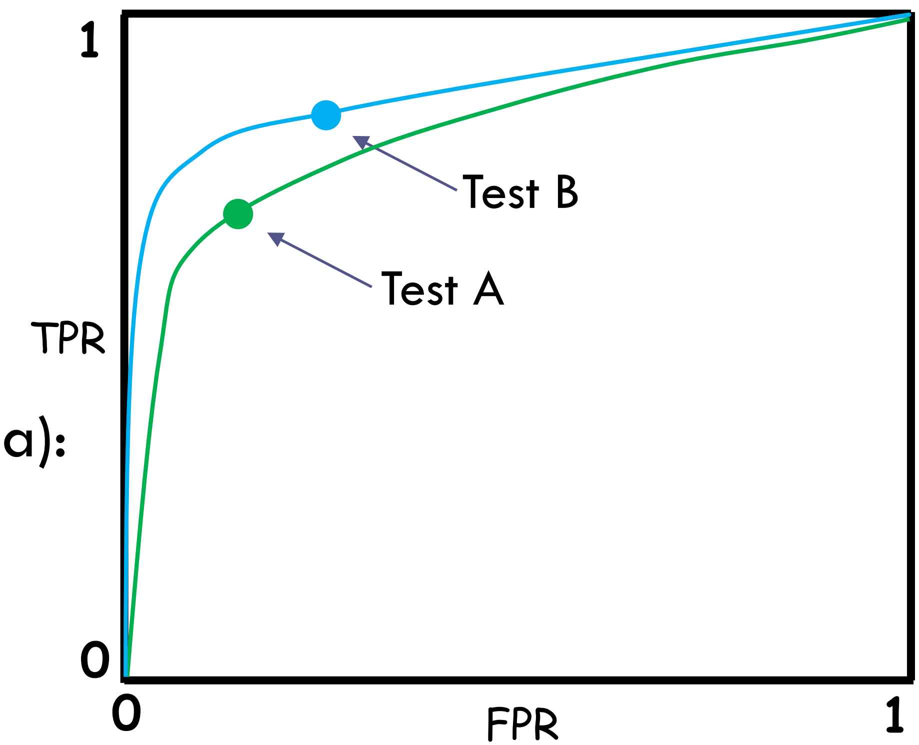


CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Sensitivity and specificity don't always make it clear which of two diagnostic tests is better: use of AUC

Which test is better?
(different ROC for each test)

Test B is better (greater ROC área):
- Higher TPR at same FPR, or
- Lower FPR at same TPR

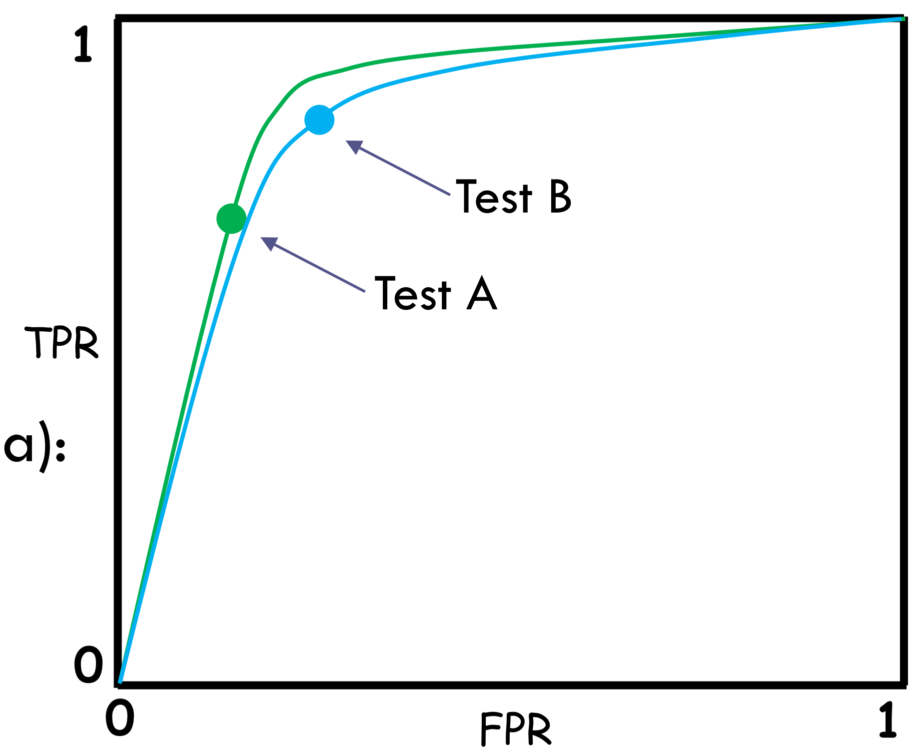


CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Sensitivity and specificity don't always make it clear which of two diagnostic tests is better: use of AUC

Which test is better?
(different ROC for each test)

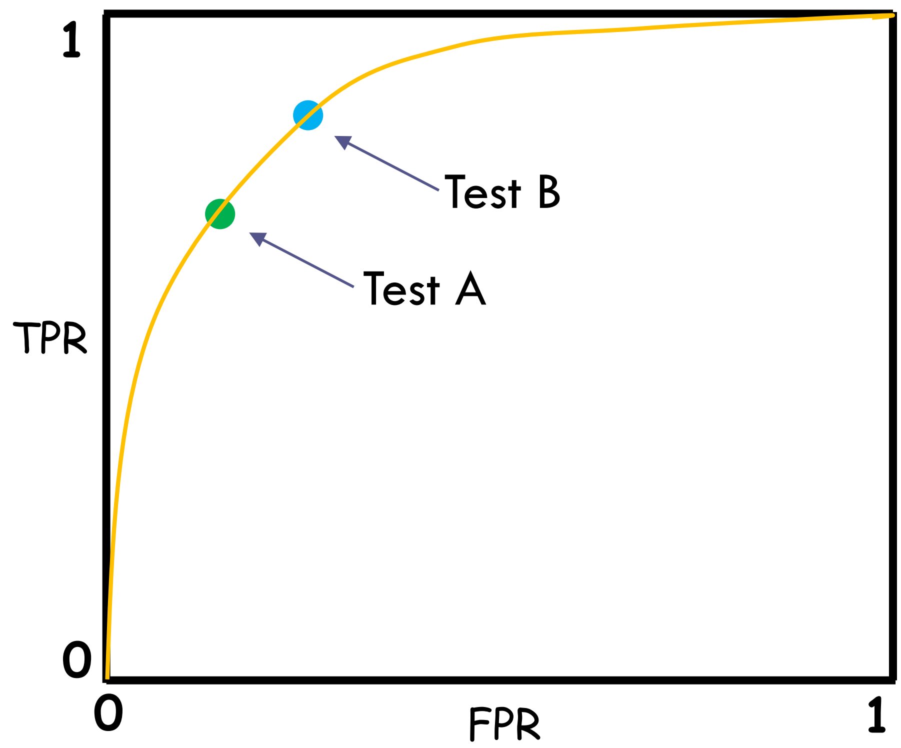
Test A is better (greater ROC área):
- Higher TPR at same FPR, or
- Lower FPR at same TPR



CAD: Computer-Aided Diagnosis

- Classifier performance: ROC Analysis
  - Sensitivity and specificity don't always make it clear which of two diagnostic tests is better: use of AUC

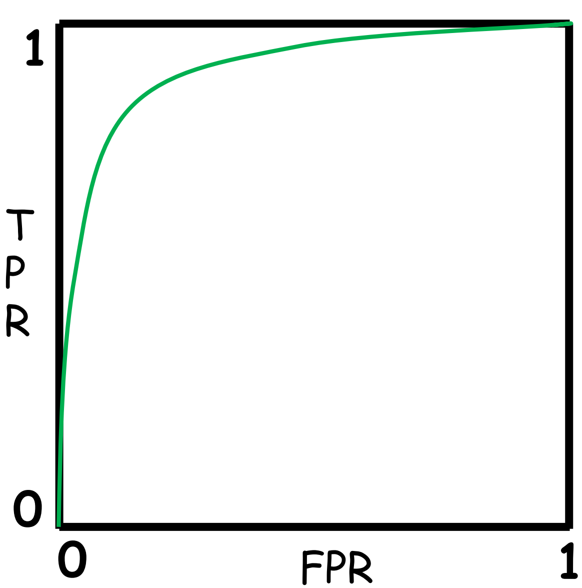
Different scenario: both tests are on the same ROC



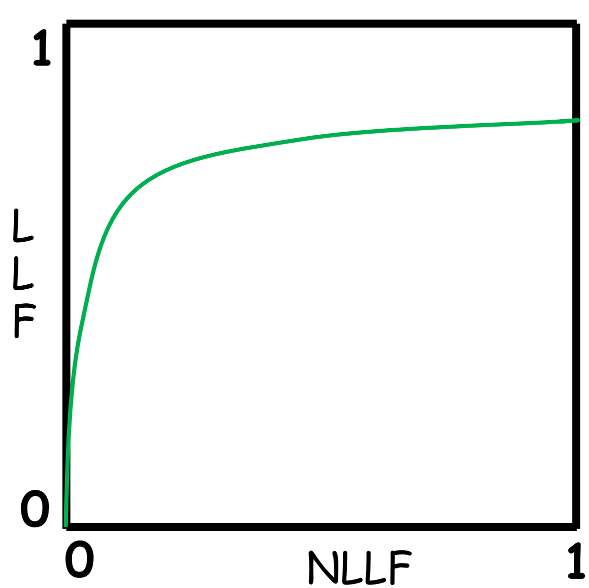
CAD: Computer-Aided Diagnosis

- Classifier performance: Generalized ROC analysis
  - Localization ROC (LROC) analysis: images must have 0 or 1 lesions, and curve depends on location-error tolerance.
  - Free-response ROC (FROC) analysis: images can have any number of lesions, and curve depends on location-error tolerance.

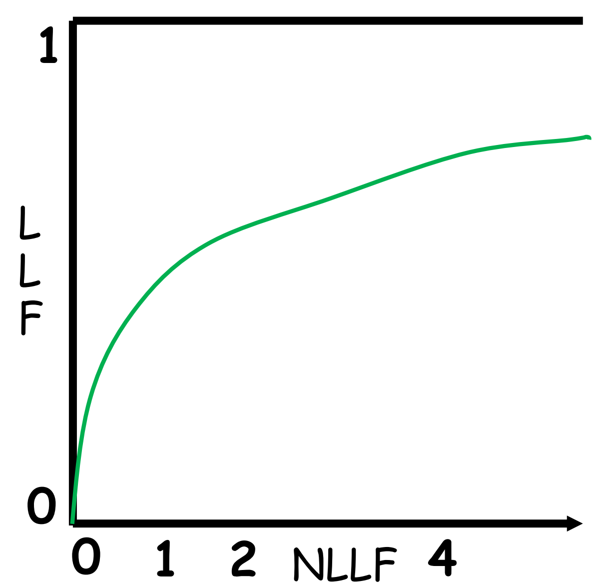
Conventional ROC curves



LROC curves



FROC curves



CAD: Computer-Aided Diagnosis

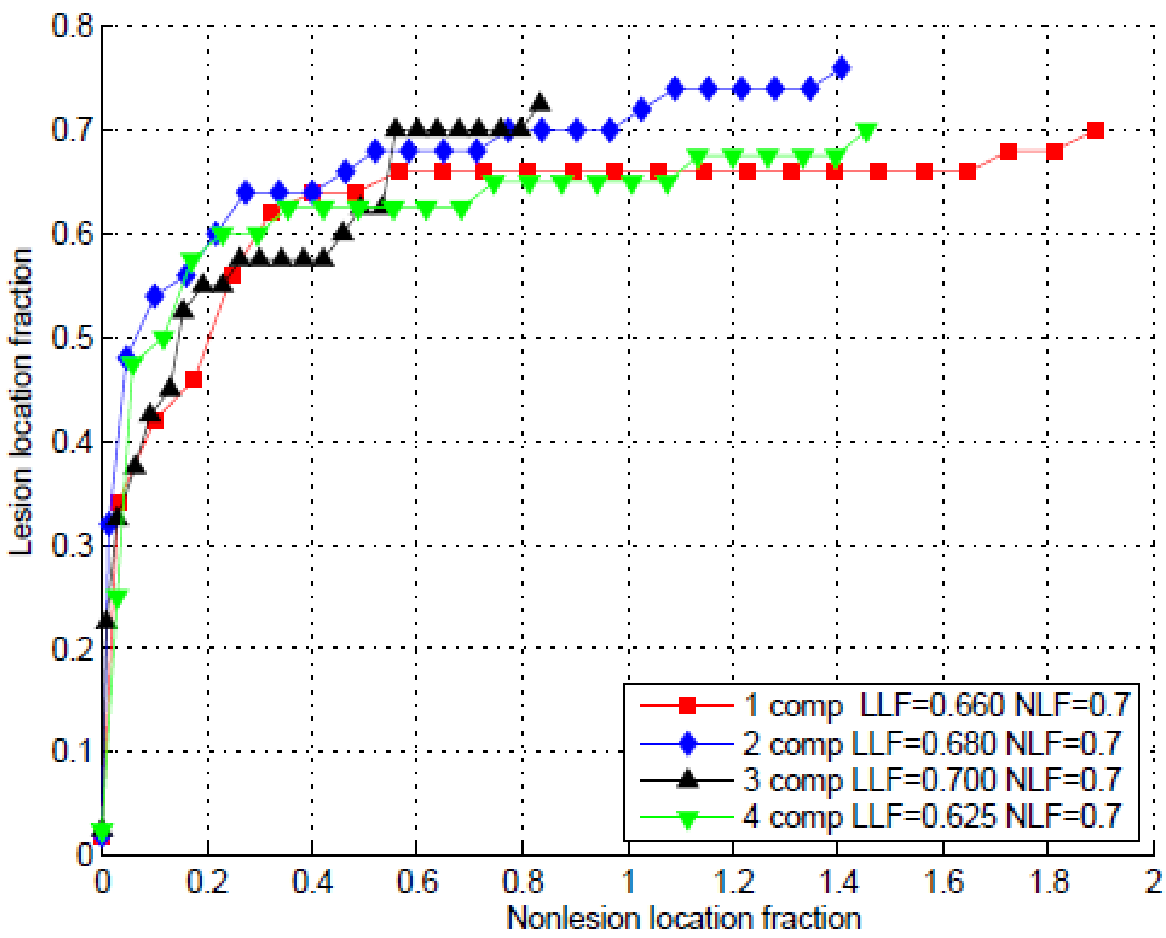
- Classifier performance: the free-response task
  - In some diagnostic examinations the truth is multi-focal (one or more diseased regions –lesions– in the patient image).
  - The truth-response table is more complex than the 2x2 ROC table.
    - In the free-response paradigm the radiologist does not know a priori how many lesions may be present in an image, and therefore must search the image for lesions and mark regions that are suspicious for disease.
    - The basic unit of data is a mark-rating pair:
      - The mark is the indicated location of the suspicious region.
      - The rating is the radiologist's estimate of the probability of disease, or confidence level (e.g. from 1 to 6, depenging on the probability of disease).

CAD: Computer-Aided Diagnosis

- Classifier performance: the free-response task
  - Some definitions:
    - LL: lesion localization (a lesion marked)
    - NL: non-lesion localization (the mark is not close to any lesion)
    - LLF: lesion localization fraction (#LL divided by total number of lesions) $0 \leq LLF \leq 1$
    - NLF: non-lesion localization fraction (there is no upper bound)
      $0 \leq NLF$
    - The FROC (**Free-Response ROC**) curve is the plot of LLF vs. NLF
      - It shows the relationship between fraction of lesions detected and number of false-positive reports per image
      - It can extend indefinitely to the right

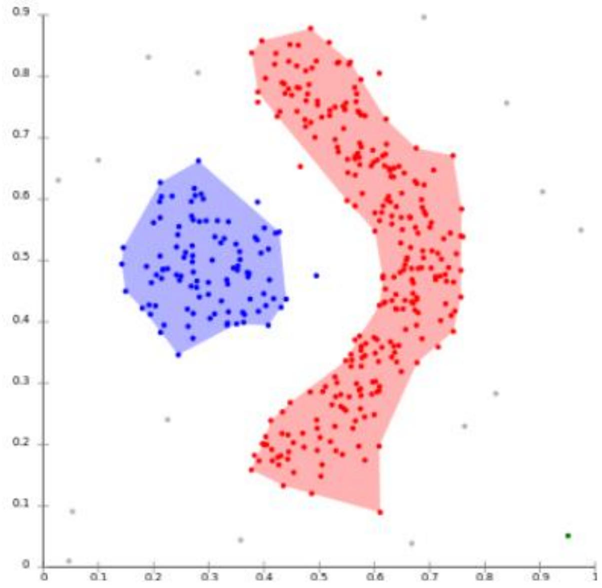
CAD: Computer-Aided Diagnosis

- Classifier performance: the free-response task



CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Clustering: the process of partitioning a set of objects into different subsets such that the data in each subset are similar to each other.
  - An important step in any clustering is to select a distance measure, which will determine how similarity of two elements is calculated.



CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - Jaccard Index
    - Dice Index
    - Cosine Index
    - Euclidean Distance
    - Manhattan Distance
    - Hamming Distance
    - Hausdorff Distance

CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used area measures:
    - **Jaccard Index** is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

    - The Jaccard distance measures dissimilarity between sample sets, and is complementary to the Jaccard Index:

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used area measures:
    - **Dice Index**, also know as Dice Similarity Coefficient (DSC) is a similarity measure over sets:

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

    - The function ranges between zero and one, like Jaccard.

CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - **<u>Cosine Index</u>** is a measure of similarity between two vectors of n dimensions by finding the angle between them, often used to compare documents in text mining.
    - Given two vectors of attributes, A and B, the cosine similarity is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

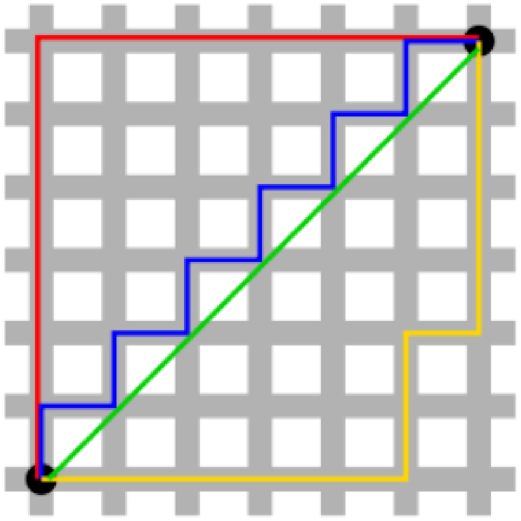
CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - **Euclidean Distance**: the Euclidean distance between points ***p*** and ***q*** is the length of the line segment connecting them.
    - In Cartesian coordinates, if $p=(p_1, p_2, \ldots, p_n)$ and $q=(q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

CAD: Computer-Aided Diagnosis

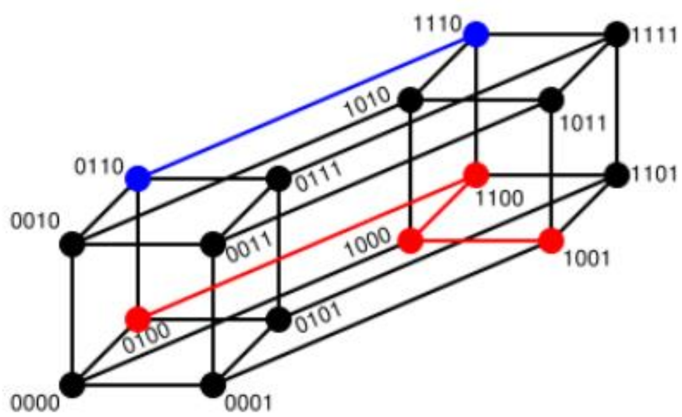
- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - **Manhattan Distance** (also known as taxicab metric, rectilinear distance, L1 distance, or city block distance) between points ***p*** and ***q*** in an n-dimensional space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segments between the points onto the coordinate axes.
    - If $p=(p_1, p_2, \ldots, p_n)$ and $q=(q_1, q_2, \ldots, q_n)$



$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

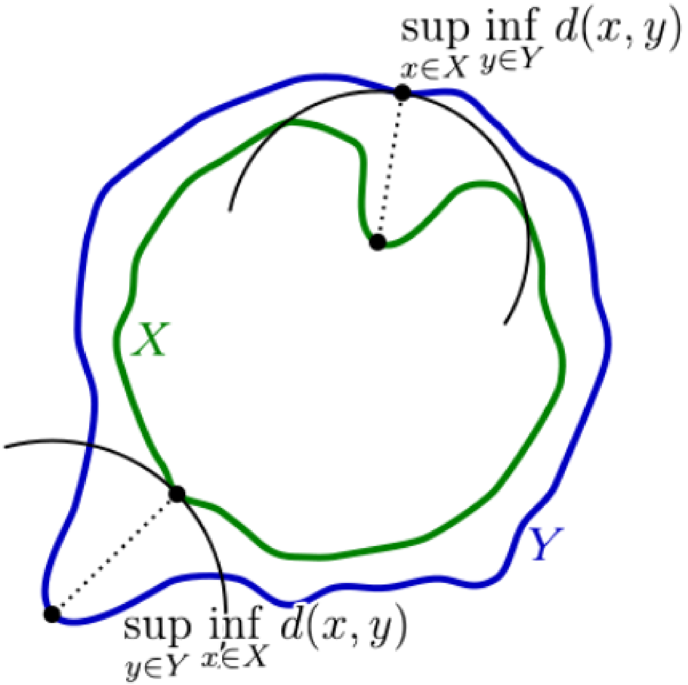
CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - **<u>Hamming Distance</u>** between two strings of equal length is the number of positions for which the corresponding symbols are different.
    - The Hamming distance can be interpreted as the number of bits which need to be changed to turn one string into other. Hamming distance can be seen as Manhattan distance between bit vectors.



CAD: Computer-Aided Diagnosis

- Classifier performance: measuring clusters distance
  - Some common used distance measures:
    - **<u>Hausdorff Distance</u>** between two sets gives an idea about their dissimilarity.
    - Two sets are close in the Hausdorff distance if every point of either set is close to some point of the other set.
    - The Hausdorff distance is the greatest of all the distances from a point in one set to the closest point in the other set.



$$d_{\mathrm{H}}(X,Y) = \max\{\,\sup_{x\in X} \inf_{y\in Y} d(x,y),\ \sup_{y\in Y} \inf_{x\in X} d(x,y)\,\}$$

References

- "An Introduction to ROC analysis". Tom Fawcett. Pattern Recognition Letters 27 (2006), 861-874.

- "Receiver Operating Characteristic Analysis: a Tool for the Quantitative Evaluation of Observer Perfomance and Imaging Systems". Charles E. Metz. J AM Coll Radiol 2006: 3:413-422

- "A comparative Study on Distance Measuring Approaches for Clustering". Shradda Pandit, Suchita Gupta. International Journal of Research in Computer Science, Volume 2, Issue 1 (2011), pp. 29-31.

- "Influence of Computer-Aided Detection on Performance of Screening Mammography". Joshua J. Fenton et al. The New England Journal of Medicine, Vol. 356 no. 14 (2007), pp. 1399-1409.

MIA – Evaluating your algorithm

Questions?