

Machine Learning – I

Richa Singh

Google classroom code: wgzuohn

Slides are prepared from several information sources on the web and books

Recap of Lecture 1

- Machine learning:
 - Supervised learning
 - Unsupervised learning
 - Reinforcement learning
- Evaluation metrics

Evaluation Metrics

- Classification: Accuracy
- Regression: Mean Squared Error
- Retrieval: Precision/Recall, F-Score
- Ranking: mean Average Precision

Sensitivity, Specificity

In a group of 500 people, 250 are known to have prostatic cancer. When tested, 200 of them had a positive test. In the people without prostatic cancer, 20 people had a positive test. What is the sensitivity and specificity of the test?

Sensitivity measures how often a test correctly generates a positive result for people who have the condition that's being tested for (also known as the "true positive" rate). A test that's highly sensitive will flag almost everyone who has the disease and not generate many false-negative results.

Specificity measures a test's ability to correctly generate a *negative* result for people who *don't* have the condition that's being tested for (also known as the "true negative" rate). A high-specificity test will correctly rule out almost everyone who *doesn't* have the disease and won't generate many false-positive results.

<https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/>

Sensitivity, Specificity

In a group of 500 people, 250 are known to have prostatic cancer. When tested, 200 of them had a positive test. In the people without prostatic cancer, 20 people had a positive test. What is the sensitivity and specificity of the test?

Answer: $200/250 = 0.8$ (sensitivity); $230/250 = 0.92$ (specificity)

Sensitivity measures how often a test correctly generates a positive result for people who have the condition that's being tested for (also known as the "true positive" rate). A test that's highly sensitive will flag almost everyone who has the disease and not generate many false-negative results.

Specificity measures a test's ability to correctly generate a *negative* result for people who *don't* have the condition that's being tested for (also known as the "true negative" rate). A high-specificity test will correctly rule out almost everyone who *doesn't* have the disease and won't generate many false-positive results.

(
<https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/>)

Precision, Recall

Example:

A search engine stores 60 documents relating to Topic A. When a search query relating to Topic A is entered, the result returns 40 documents relating to Topic A, and 10 documents related to some other topic. What is the precision and recall for the given search?

Precision, Recall

Example:

A search engine stores 60 documents relating to Topic A. When a search query relating to Topic A is entered, the result returns 40 documents relating to Topic A, and 10 documents related to some other topic. What is the precision and recall for the given search?

Answer: $40/50 = 0.8$ (precision); $40/60 = 0.66$ (recall)

Type-I error, Type-II error

A binary classifier classifies given objects into two classes - bags (+ve class) and purses (-ve class). Out of 90 bags and 20 purses, it classified 80 bags correctly and none of the purses. What is the confusion matrix? What is the Type-1, Type-II error?

- Answer:
- TP-80; TN-0; FP-20; FN-10
- Type I: $20/20 = 1$
- Type II: $10/90 = 0.11$

Solve the questions shared via WebEx chat

- Clustering: Normalized Mutual Information

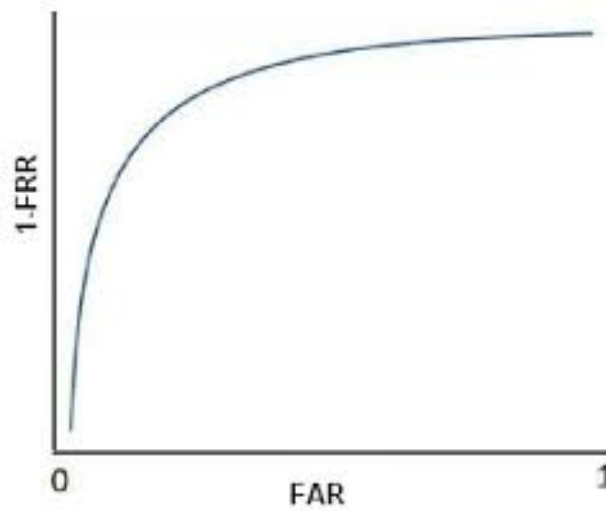
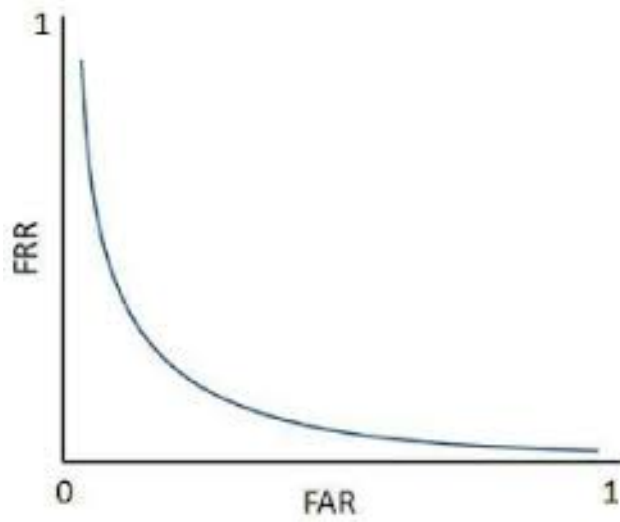
$$NMI(Y, C) = \frac{2 \times I(Y; C)}{[H(Y) + H(C)]}$$

- Y = class labels
- C = cluster labels
- $H(.)$ = Entropy
- $I(Y; C)$ = Mutual Information b/w Y and C
- Note: All logs are base-2.

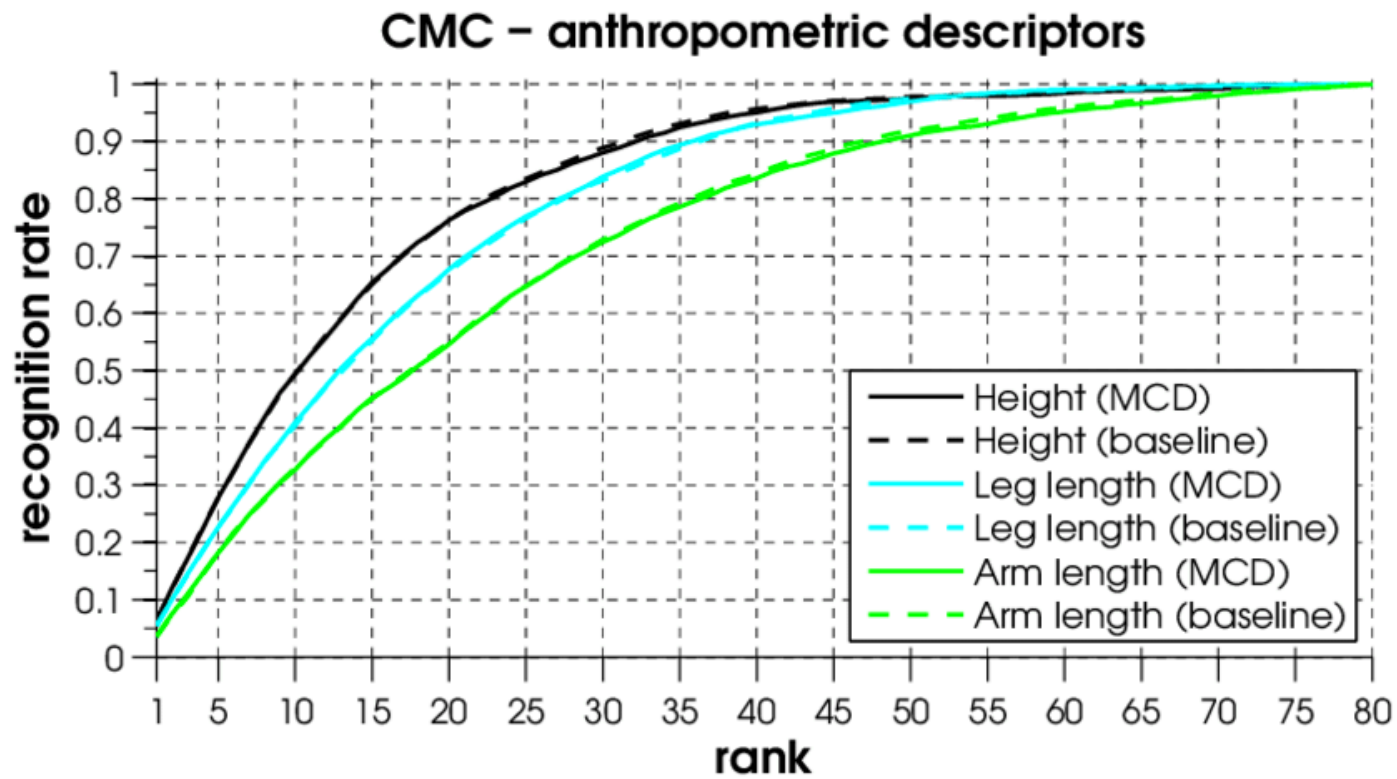
Performance Evaluation

- Receiver operating characteristics (ROC) curve
 - For authentication/verification
 - False positive rate vs true positive rate
- Detection error-tradeoff (DET) curve
 - False positive rate vs false negative rate
- Cumulative match curve (CMC)
 - Rank vs identification accuracy

ROC Curve







CMC Curve

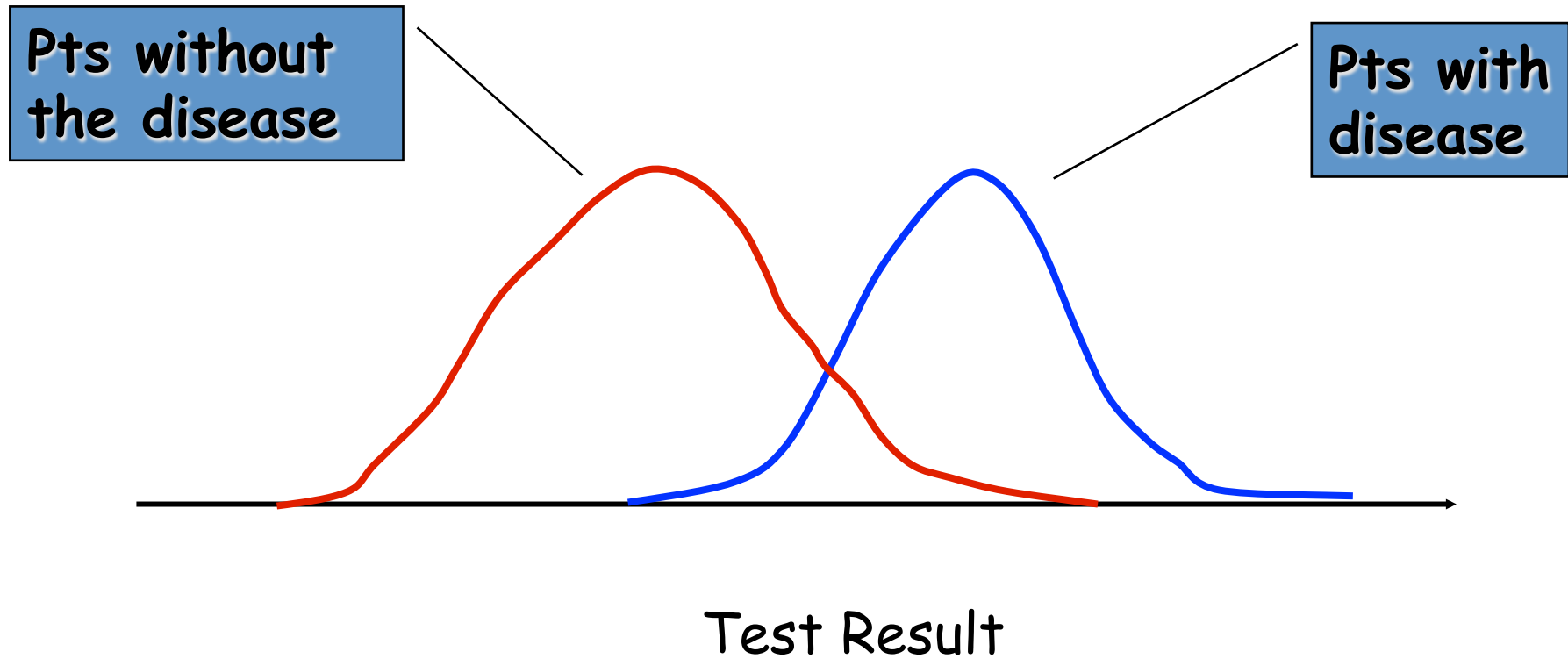


How to draw ROC curve?

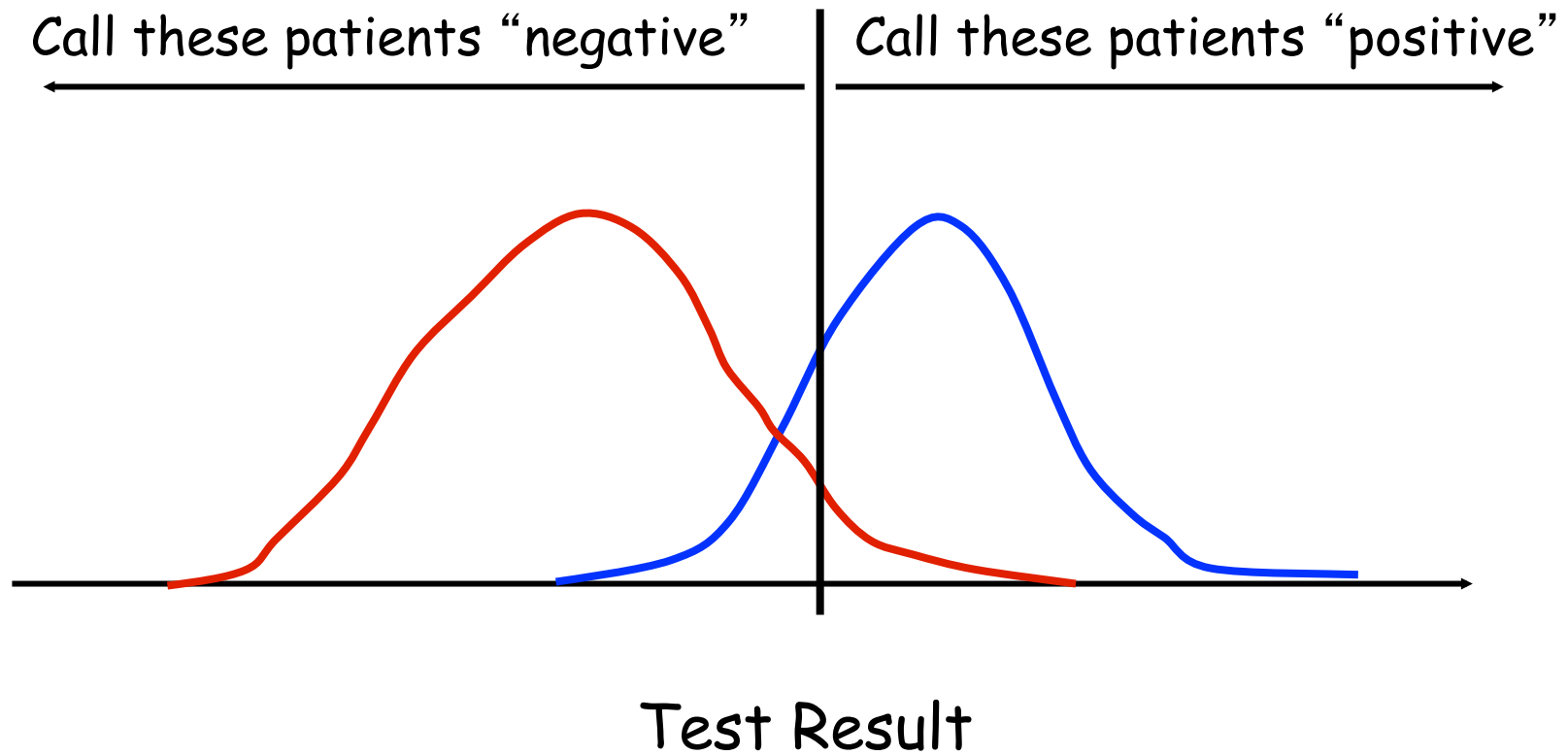
True disease state vs. Test result

<div>Disease \ Test</div>	not rejected	rejected
No disease (D = 0)	 specificity	 Type I error (False +) α
Disease (D = 1)	 Type II error (False -) β	 Power $1 - \beta$; sensitivity

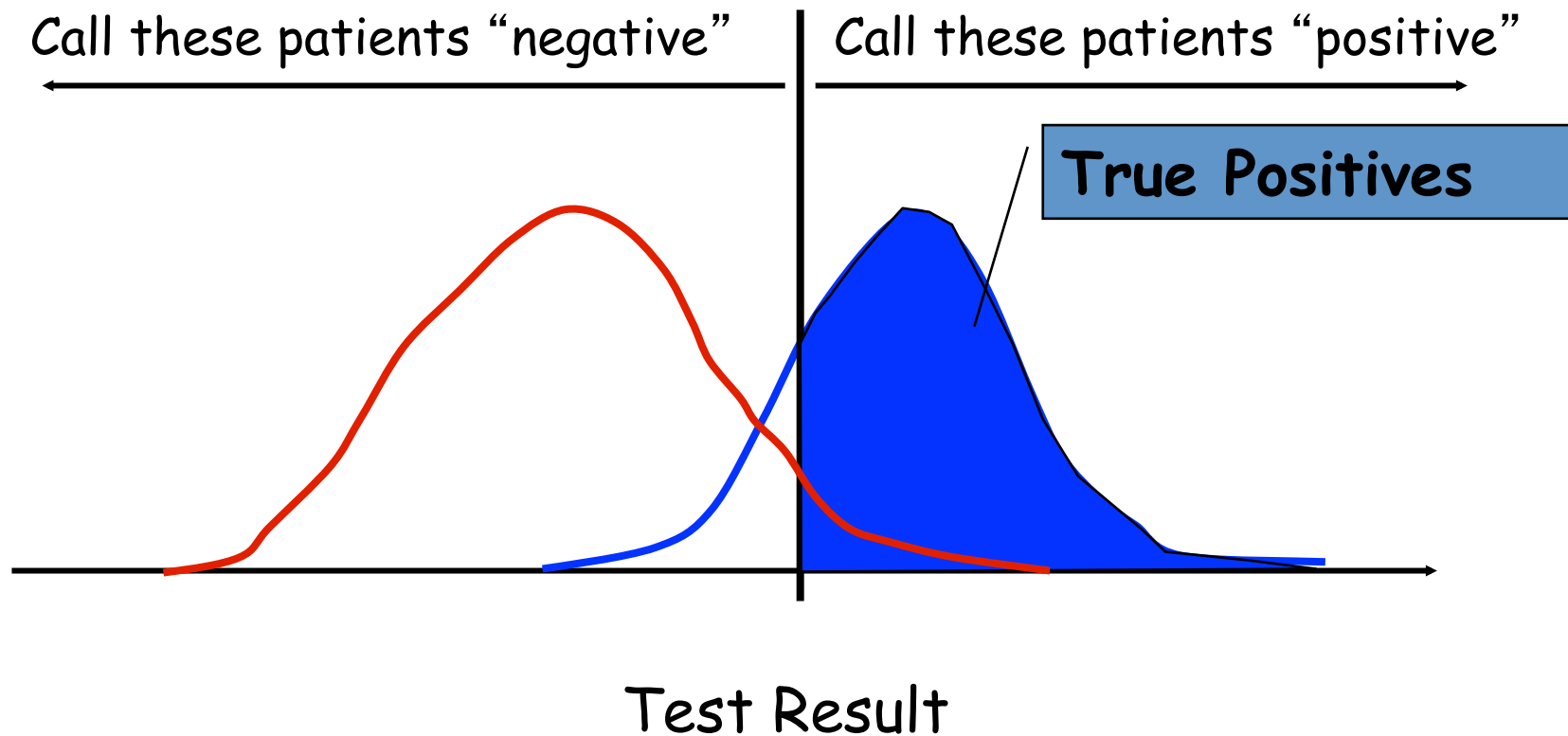
Specific Example



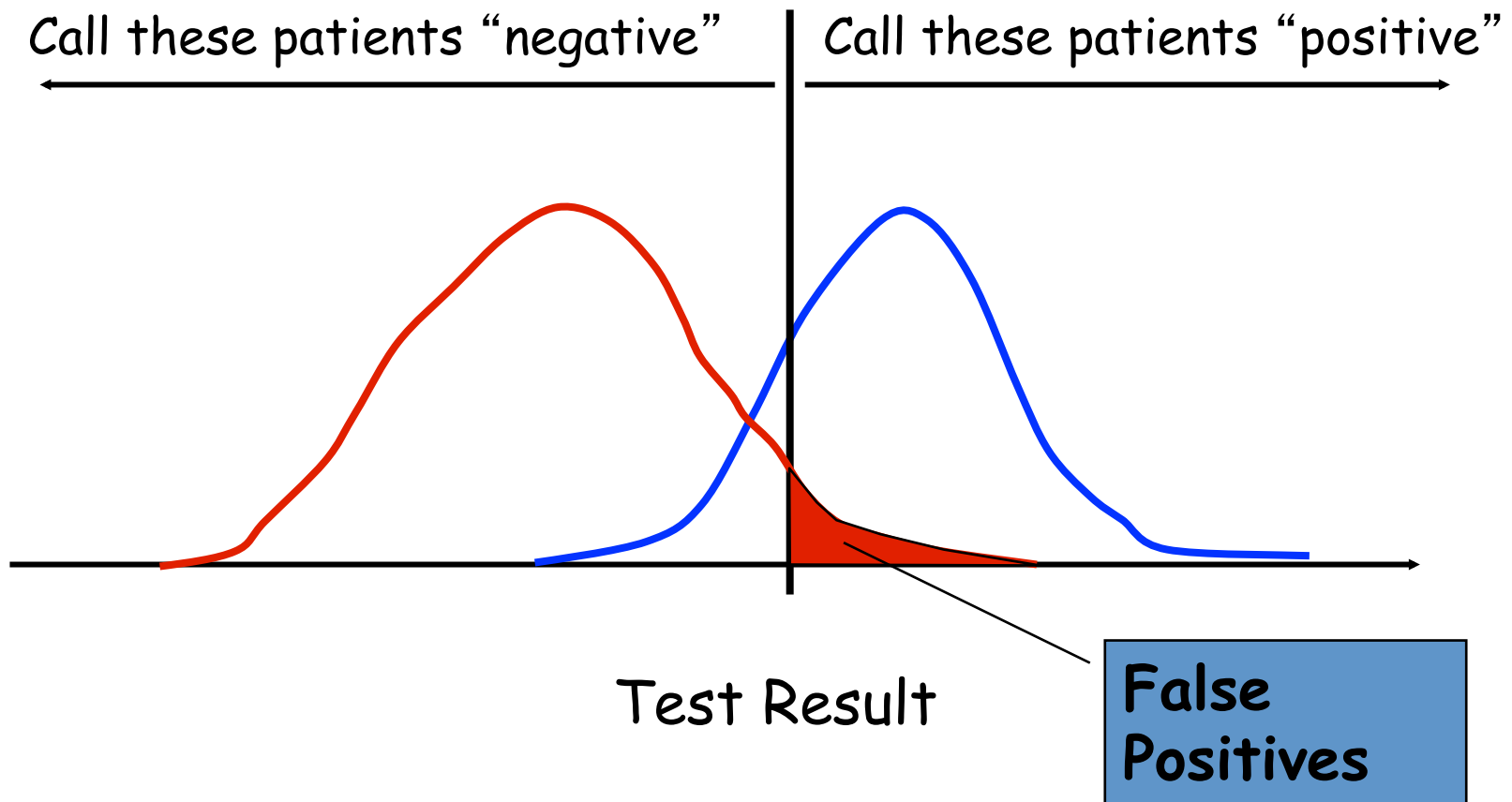
Threshold



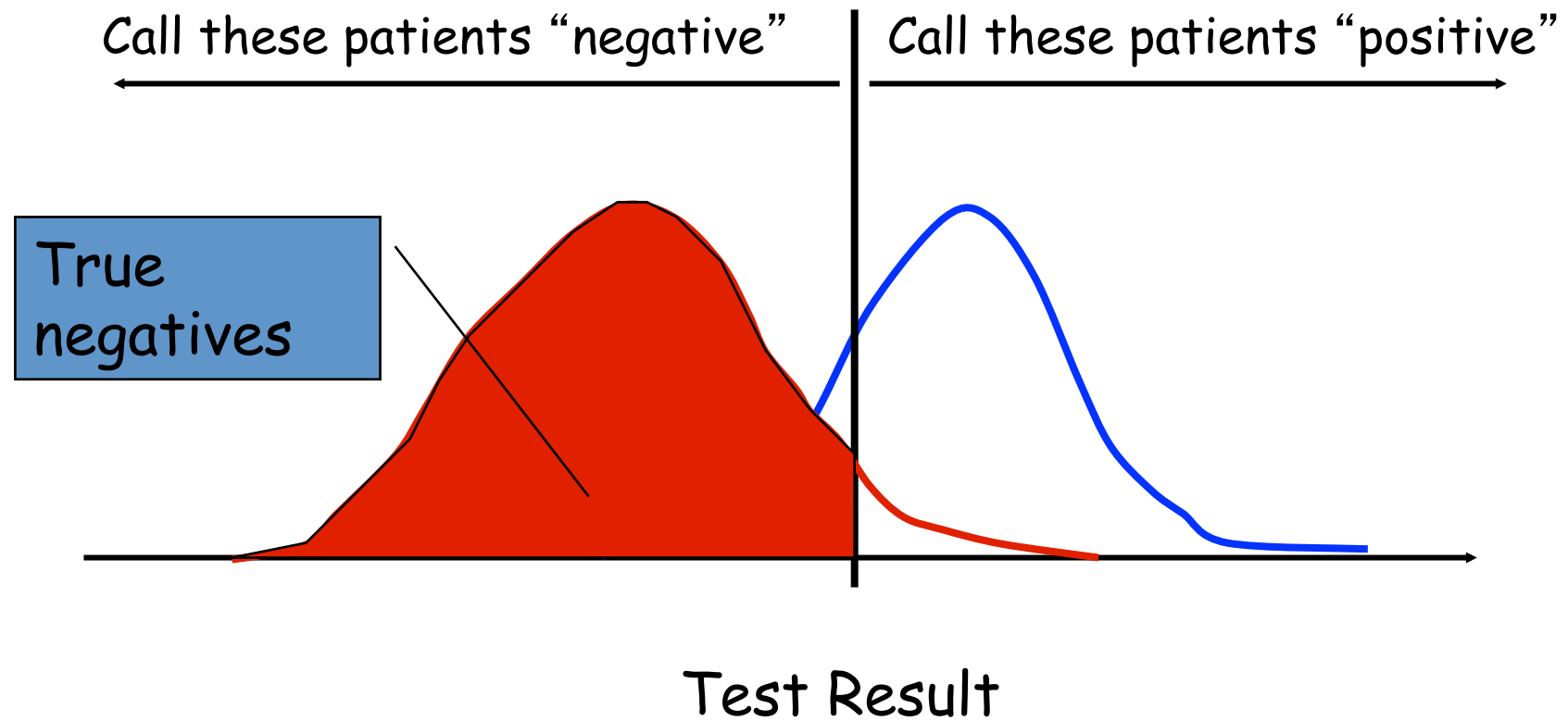
Some definitions ...



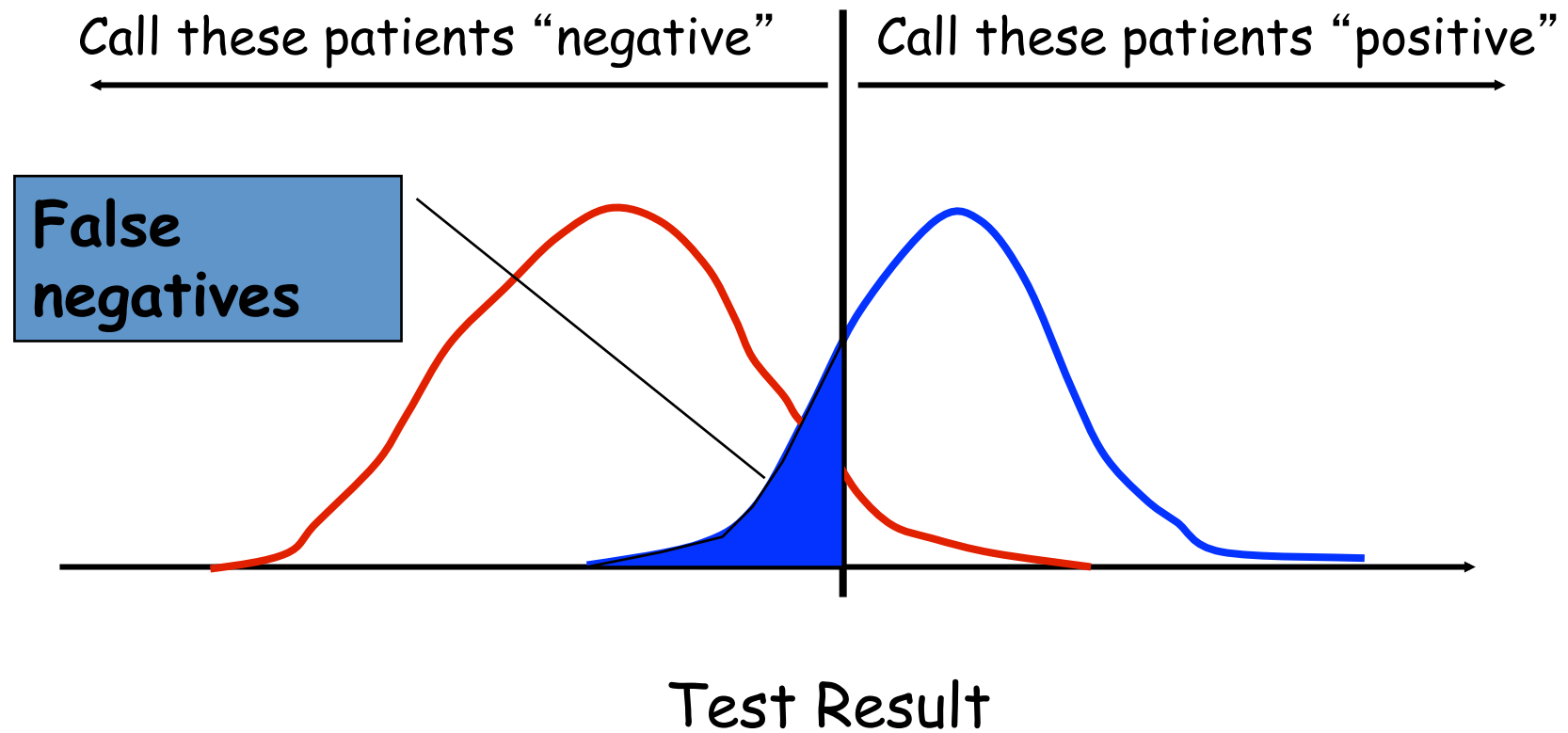
without the disease
with the disease



without the disease
with the disease

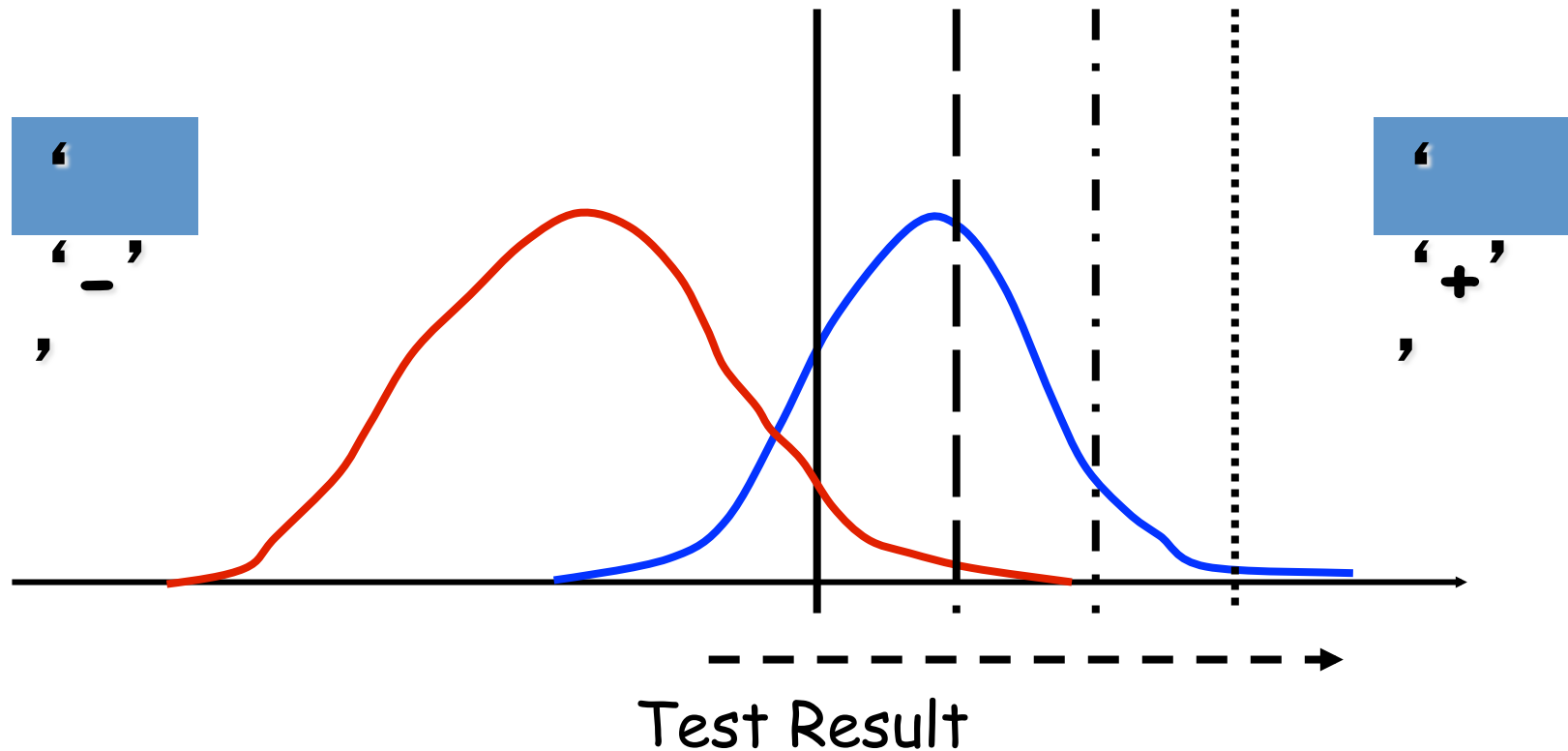


without the disease
with the disease



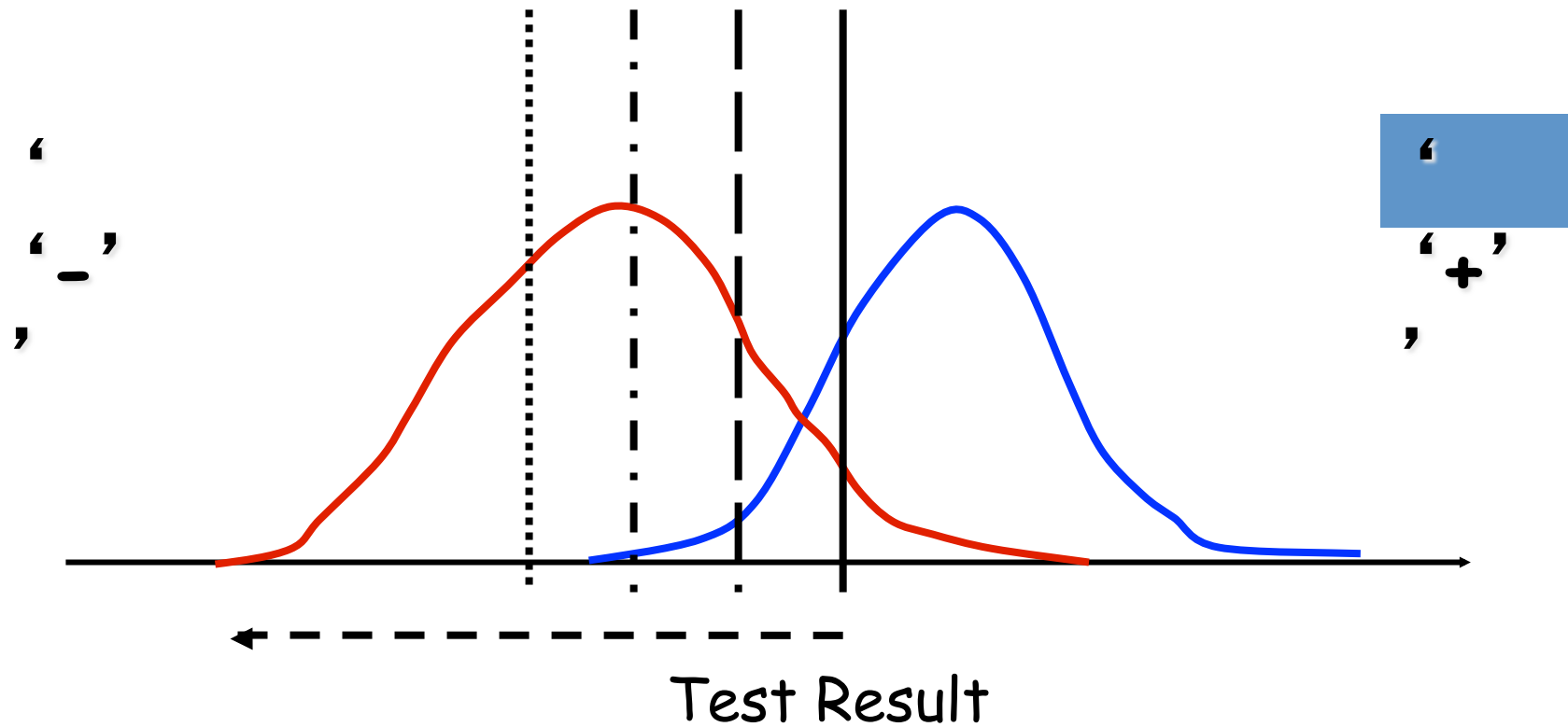
without the disease
with the disease

Moving the Threshold: right



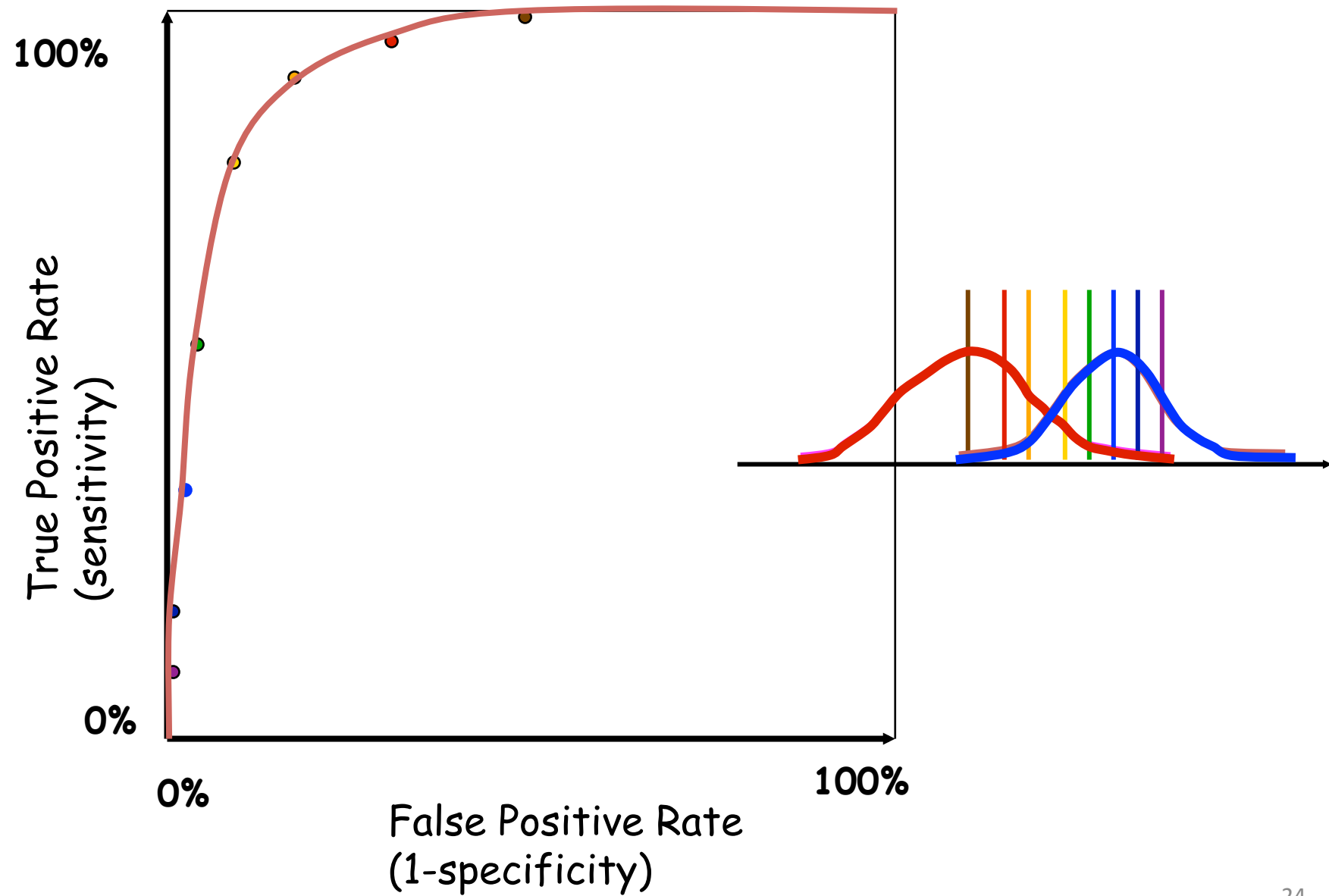
without the disease
with the disease

Moving the Threshold: left



without the disease
with the disease

ROC curve

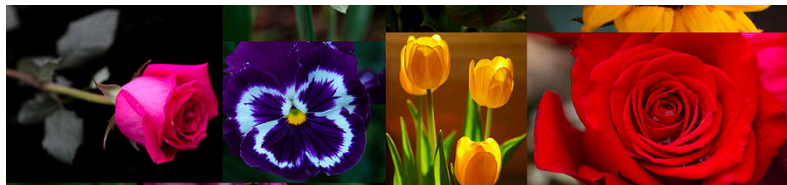


How to draw CMC curve?

Samples in the database



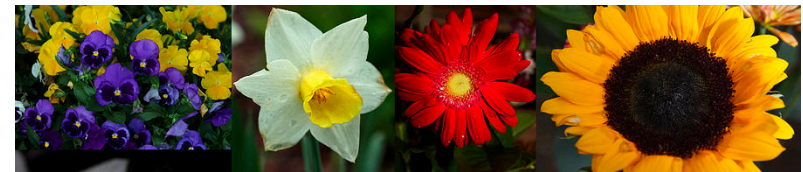
Query



Top-4 retrieved images

Correct answer at rank - 1

Query



Top-4 retrieved images

Correct answer at rank - 4

How to compute rank accuracies

- Rank-1 accuracy: 50%
- Rank-2 accuracy: 50 + 0 %
- Rank-3 accuracy: 50 + 0 + 0 %
- Rank-4 accuracy: 50 + 0 + 0 + 50 %

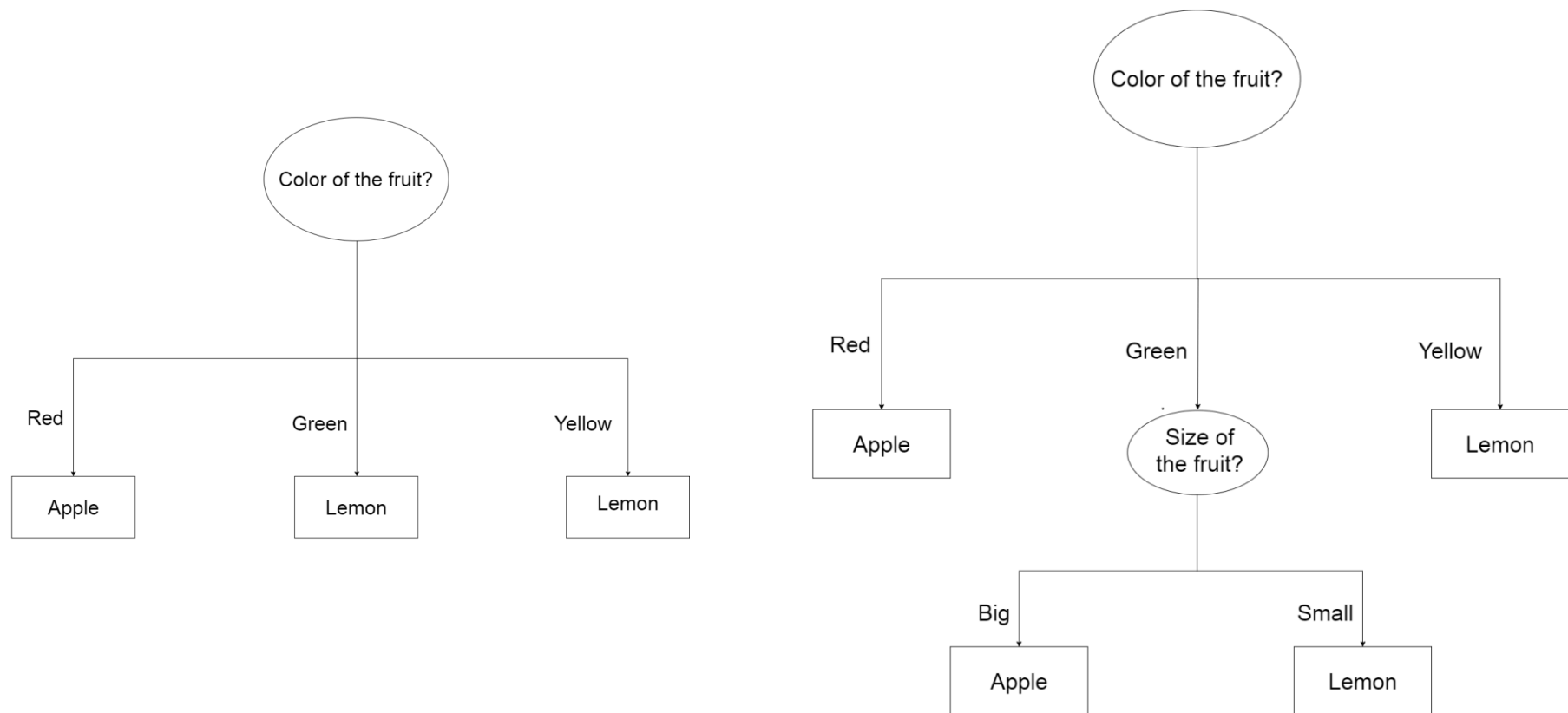
Cumulative in
nature

Non-decreasing

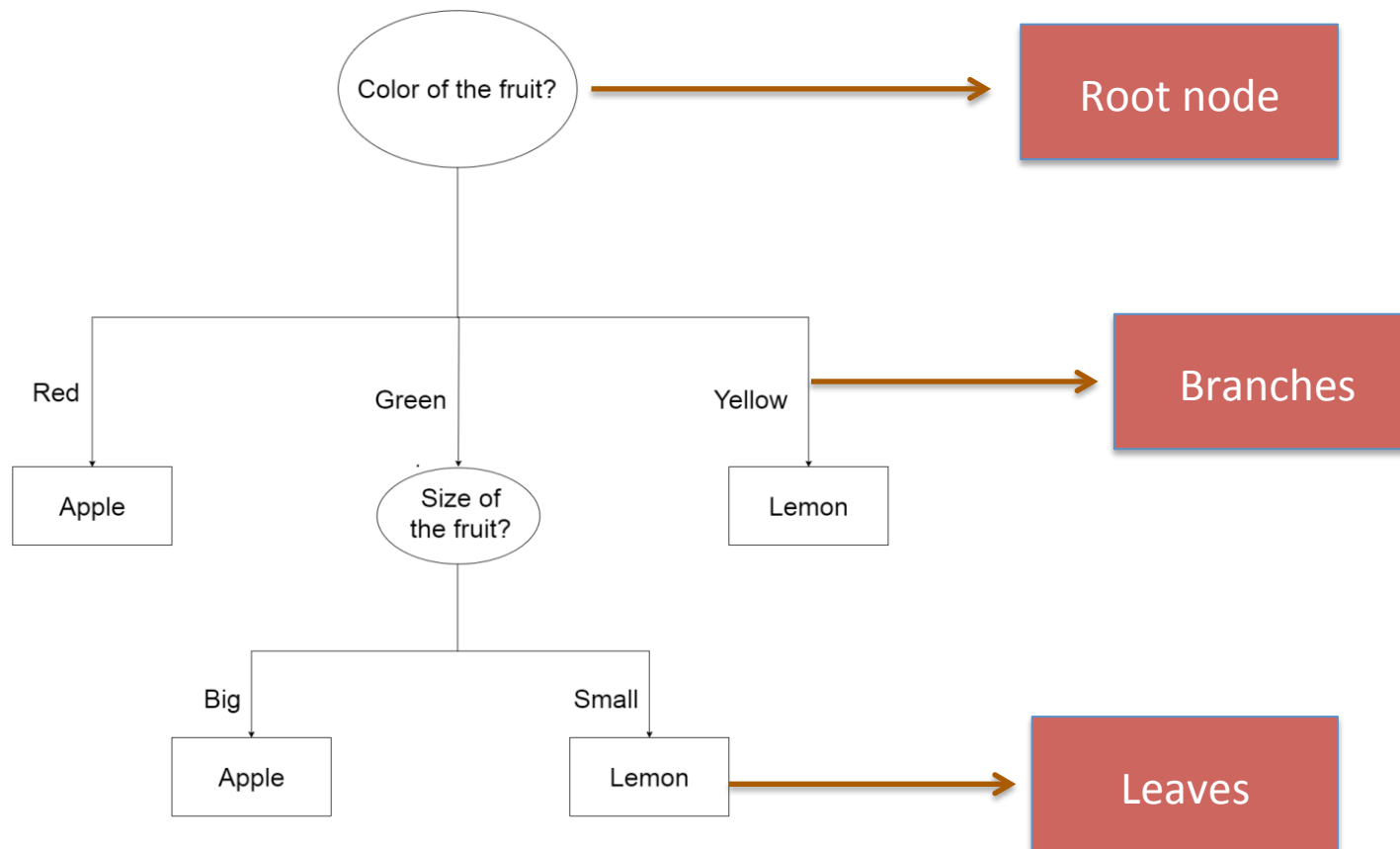
Decision Trees

Decision Trees

- Classify between lemon and oranges



Decision Trees

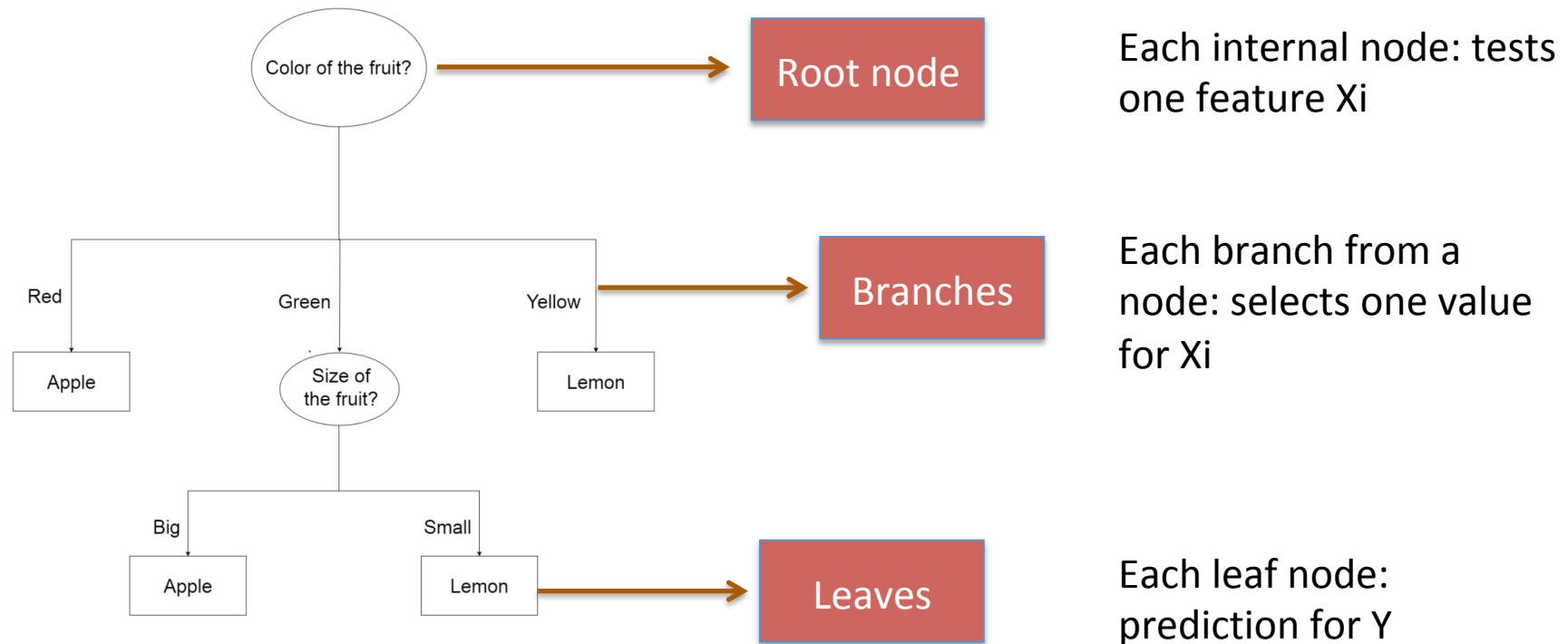


Rules for classifying data using attributes

- The tree consists of decision nodes and leaf nodes.
- A decision node has two or more branches, each representing values for the attribute tested.
- A leaf node attribute produces a homogeneous result (all in one class), which does not require additional classification testing

\mathcal{F} – Decision Trees

$$f(X_1, X_2, X_3) \in \mathcal{F}$$



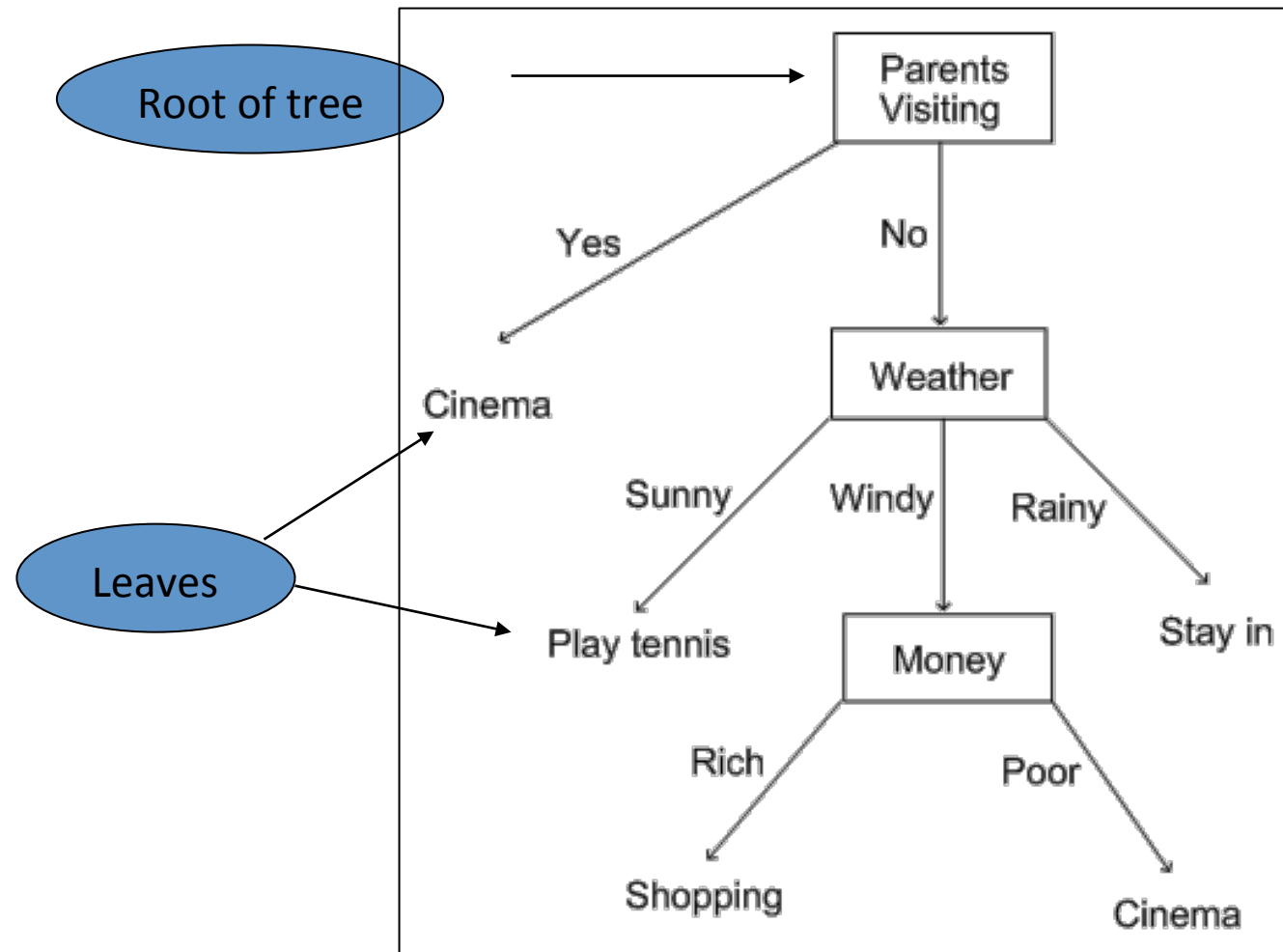
Features can be discrete, continuous or categorical

- Features can be discrete, continuous or categorical
- Each internal node: test some set of features $\{X_i\}$
- Each branch from a node: selects a set of value for $\{X_i\}$
- Each leaf node: prediction for Y

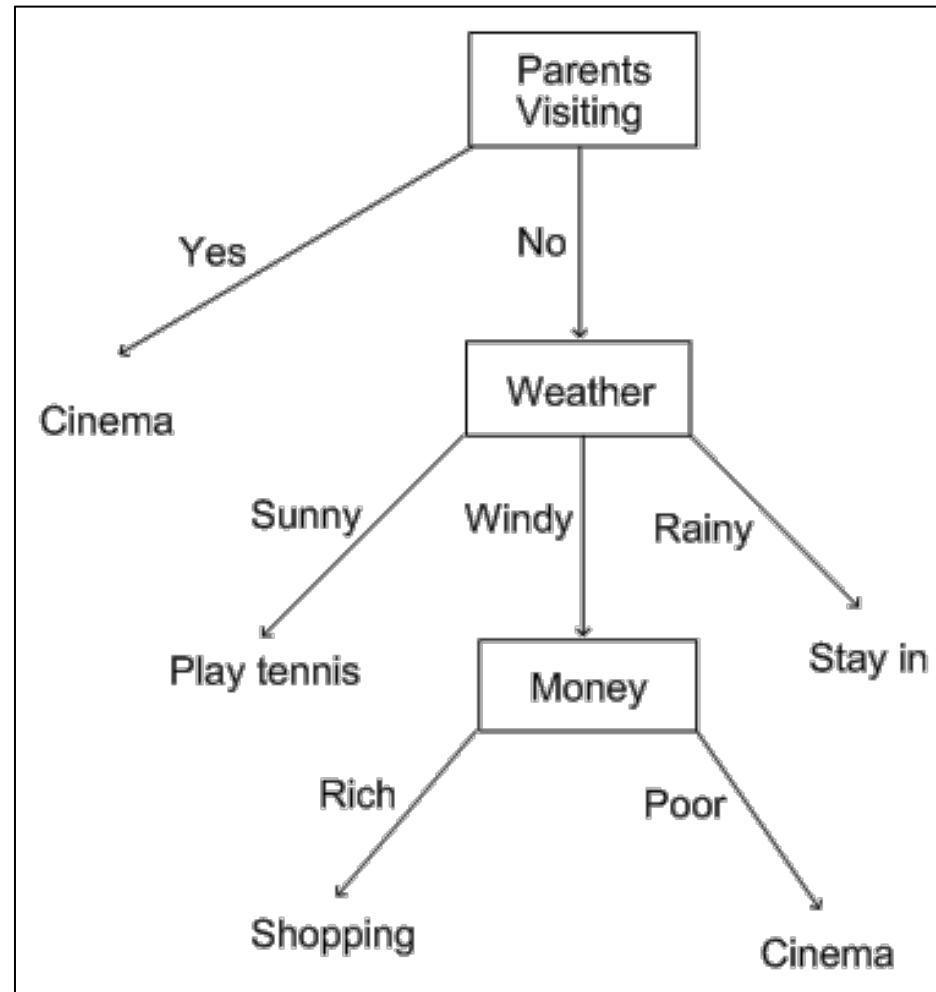
Example: What to do this Weekend?

- If my parents are visiting
 - We'll go to the cinema
- If not
 - Then, if it's sunny I'll play tennis
 - But if it's windy and I'm rich, I'll go shopping
 - If it's windy and I'm poor, I'll go to the cinema
 - If it's rainy, I'll stay in

Written as a Decision Tree

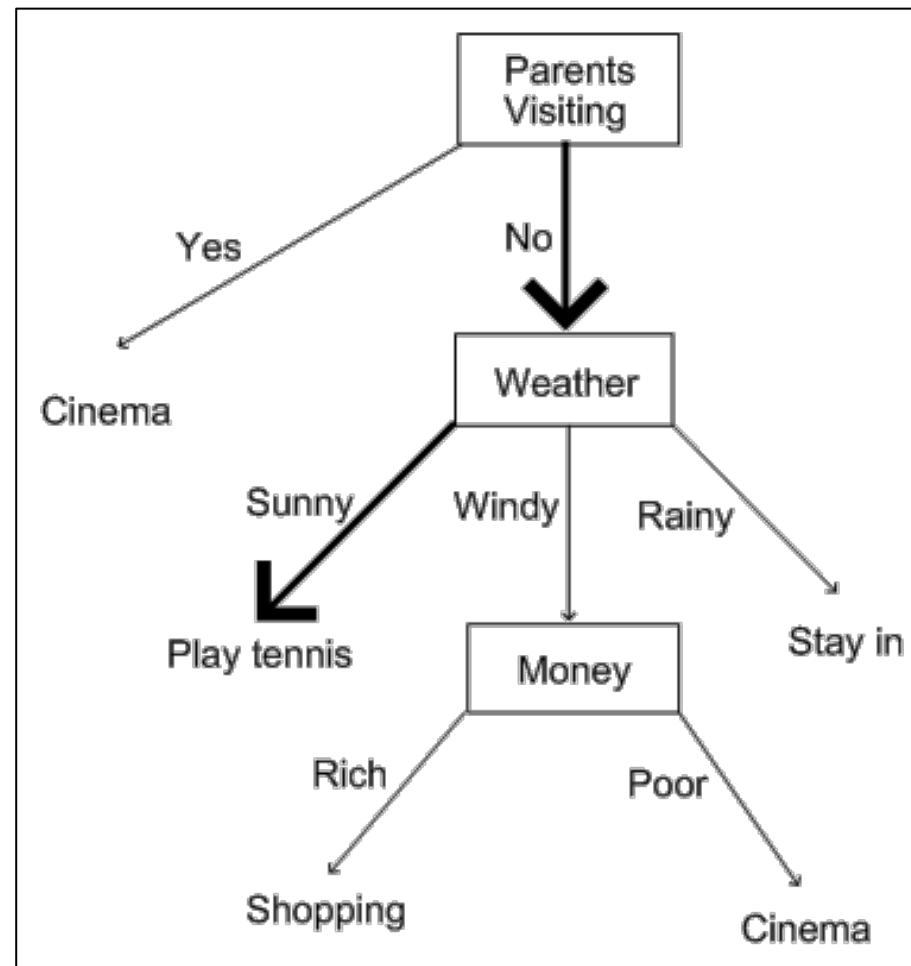


Using the Decision Tree (No parents on a Sunny Day)



Using the Decision Tree

(No parents on a Sunny Day)



From Decision Trees to Logic

- Read from the root to every tip
 - If this and this and this ... and this, then do this
- In our example:
 - If no_parents and sunny_day, then play_tennis
 - $\text{no_parents} \wedge \text{sunny_day} \rightarrow \text{play_tennis}$

How to design a decision tree

- Decision tree can be seen as rules for performing a categorisation
 - E.g., “what kind of weekend will this be?”
- Remember that we’re learning from examples
 - Not turning thought processes into decision trees
- The major question in decision tree learning is
 - Which nodes to put in which positions
 - Including the root node and the leaf nodes

What do you think: how should we compute
which nodes to put in which positions?

The ID3 Algorithm

- Invented by J. Ross Quinlan in 1979
- ID3 uses a measure called Information Gain
 - Used to choose which node to put next
- Node with the highest information gain is chosen
 - When there are no choices, a leaf node is put on
- Builds the tree from the top down, with no backtracking
- Information Gain is used to select the most useful attribute for classification

Entropy – General Idea

- From Tom Mitchell's book:
 - “In order to define information gain precisely, we begin by defining a measure commonly used in information theory, called entropy that characterizes the (im)purity of an arbitrary collection of examples”
- A notion of impurity in data
- A formula to calculate the homogeneity of a sample
- A completely homogeneous sample has entropy of 0
- An equally divided sample has entropy of 1

Entropy - Formulae

- Given a set of examples, S
- For example, in a binary categorization
 - Where p_+ is the proportion of positives
 - And p_- is the proportion of negatives

$$\textit{Entropy}(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

- For examples belonging to classes c_1 to c_n
 - Where p_n is the proportion of examples in c_n

$$\textit{Entropy}(S) \equiv \sum_{i=1}^n -p_i \log_2 p_i$$

Entropy Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Entropy Example

Entropy(S) =

$$\begin{aligned} & - (9/14) \text{Log}_2 (9/14) - (5/14) \text{Log}_2 (5/14) \\ & = 0.940 \end{aligned}$$