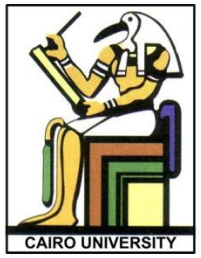
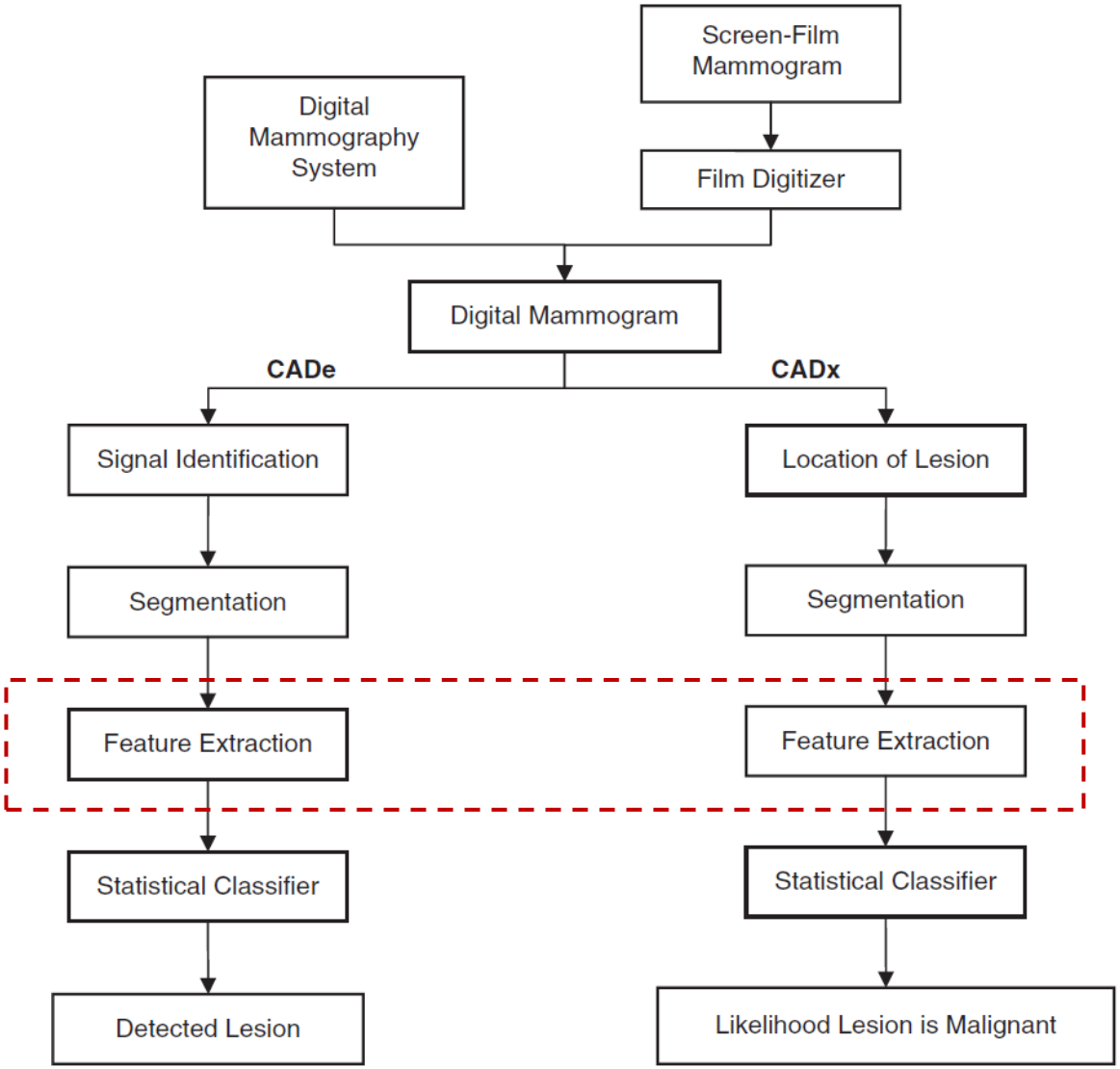


ADVANCED TOPICS IN BIOMEDICAL ENGINEERING

Topic 5: Feature Selection

Prof. Yasser Mostafa Kadah

CAD


Screen-Film Mammogram
Digital Mammography System
Film Digitizer
Digital Mammogram
CADe
CADx
Signal Identification
Location of Lesion
Segmentation
Segmentation
Feature Extraction
Feature Extraction
Statistical Classifier
Statistical Classifier
Detected Lesion
Likelihood Lesion is Malignant

Feature Extraction and Selection

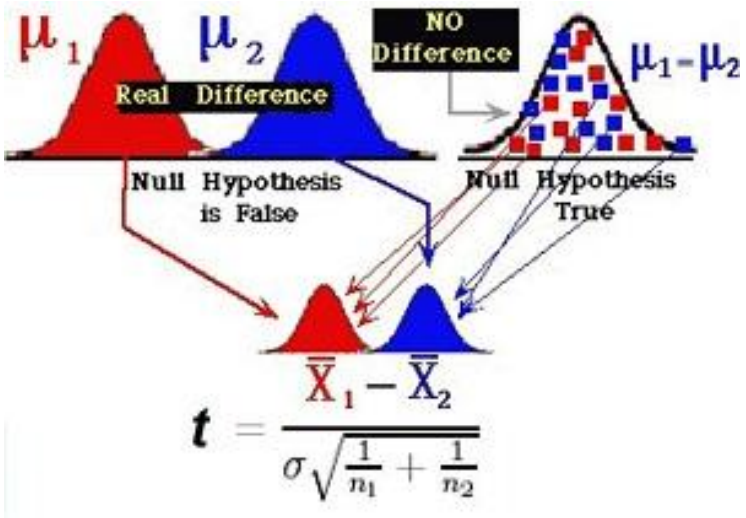
- Feature extraction and selection in pattern recognition are based on finding mathematical methods for reducing dimensionality of pattern Representation
- A lower-dimensional representation based on independent pattern descriptors is important
  - Plays crucial role in determining separating properties of pattern classes
- The choice of features, attributes, or measurements to be included has an important influence on:
  - (1) accuracy of classification
  - (2) time needed for classification
  - (3) number of examples needed for learning
  - (4) cost of performing classification

Feature Selection Goal

- Features extracted from images need not represent significant information to diagnosis
  - May describe aspects of no relevance to the pathology of interest
  - May vary a lot with acquisition settings (pose, processing, etc.)
- Several problems should be mitigated in feature selection
  - Features that do not correlate with pathology
  - Features that are not independent
- Building classifiers with features that are not properly selected will cause problems in the training phase and will not yield the best overall classification accuracy

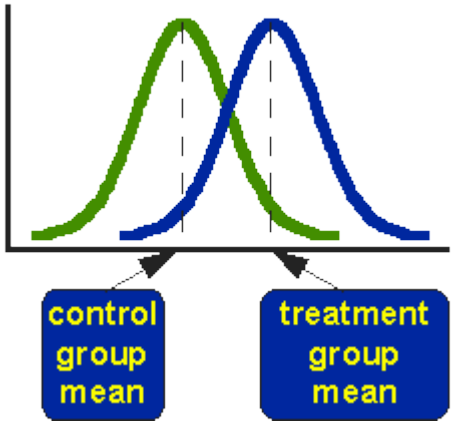
Statistical Significance of Features

- Idea: if the feature changes consistently with pathology, then the hypothesis of a statistically significant difference between the set of values for normal and abnormal cases will be true
  - Inferential Statistical tests like Student t-test should detect a difference



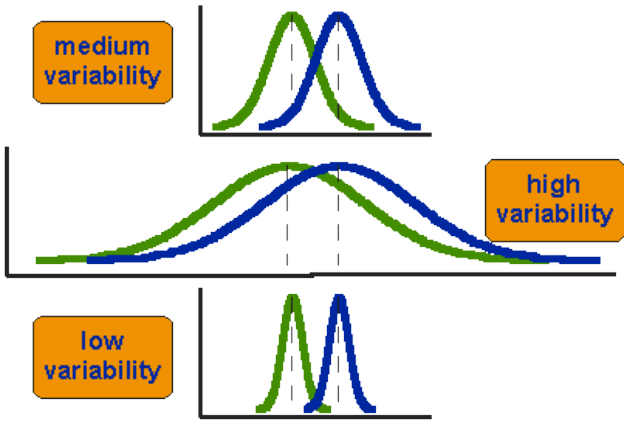
Statistical Significance of Features

- Student t-test steps:
  - Consider a particular feature of interest
  - Divide the values into two sets for normal and abnormal cases
  - Compute the mean and standard deviation for both sets
  - Use the t-test to compute the p-value of the null hypothesis that both sets do not have a statistically significant difference
  - The feature is suitable if the p-value is small (e.g., 0.05, 0.01, etc.)

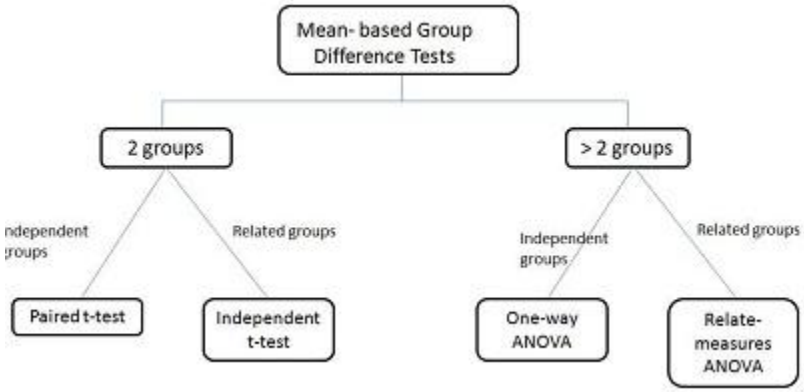


Statistical Significance of Features

- Important to keep in mind that large difference in value does not mean statistical significance
  - Data dispersion is a key factor



- General form: multiple groups
  - Diagnosis not detection
  - More general inferential statistics
- Nonparametric methods
  - Kolmogorov-Smirnov test
  - Wilcoxon signed rank test



Assessment of Feature Independence

- Some features may end up being dependent
  - Example: feature computed as a constant factor of another
  - Only one of them should be included as the input to classification stage
- Several methods can help identify such dependence
  - Pearson's linear correlation coefficient
  - Principal component analysis

Pearson's linear correlation coefficient

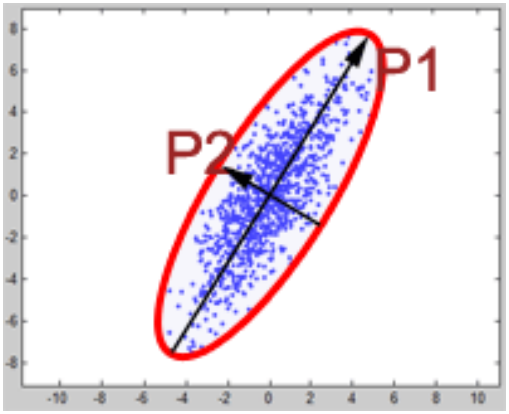
- Computes the correlation between a given pair of features
- Computing formula:

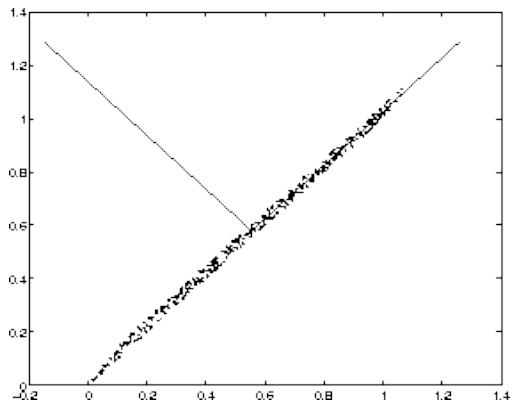
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

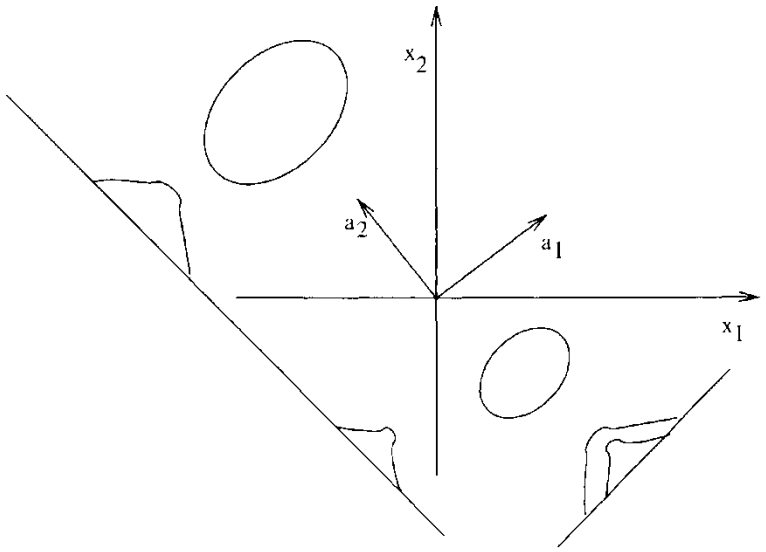
- The value of r lies between 1 and 1, inclusive
  - Value of 1 is called "complete positive correlation": straight line y = c x
  - Value of -1 is called "complete negative correlation": straight line y= -c x
  - Value of r near zero indicates that the variables x and y are *uncorrelated*

Principal Component Analysis

- Computes the eigenvalue decomposition of the matrix of features
  - Rank of matrix = number of independent features
  - Directions of principal components may have different performance in classification







Retrospective Assessment of Features

- Retrospective: evaluation after seeing the classification results based on these features
- Basic idea: use for classification and then choose the features that produce the best results
  - Exhaustive search
  - Branch and Bound Algorithm
  - Max-Min Feature Selection
  - Sequential Forward and Sequential Backward Selection
  - Fisher's Linear Discriminant

Exhaustive Search

- Let y with $y = [y_1, y_2, \ldots, y_D]$ be a pattern vector, exhaustive search selects the **d** best features out of the maximal available features D as to minimize the classification error
- The resulting number of total combinations is:

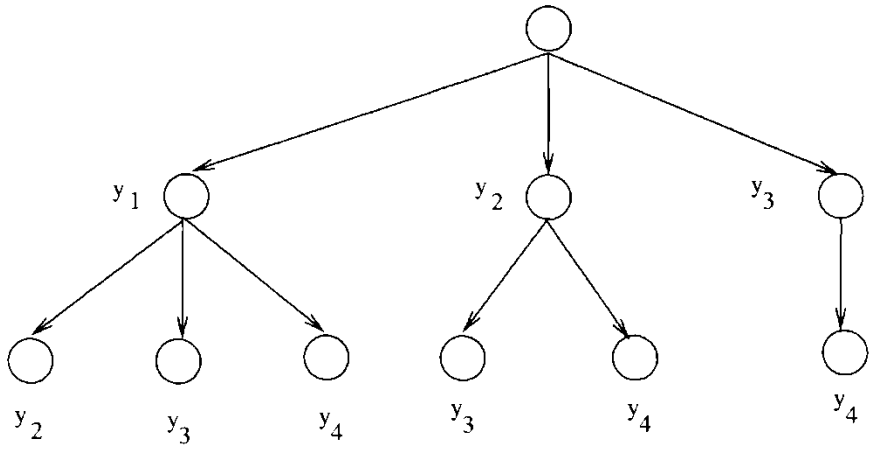
$$\binom{D}{d} = \frac{D!}{(D-d)!d!}$$

- Main advantage: guaranteed best answer
- Main disadvantage of exhaustive search is that the total number of combinations increases exponentially with the dimension of the feature vector

Branch and Bound Algorithm

- Solves the problem of huge computational cost associated with the exhaustive search
- New technique to determine the optimal feature set without explicit evaluation of all possible combinations of d features
- This technique is applicable when the separability criterion is monotonic: $\text{If} \quad \chi_1 \subset \chi_2 \subset \cdots \chi_D, \quad \text{then} \quad J(\chi_1) \leq J(\chi_2) \leq \cdots J(\chi_D)$

- Combinatorial optimization
- Efficient computation



Assignments

- Apply 2 feature selection methods to the features obtained from previous tasks and compare their results
- Add a random data vector as a feature to your extracted features and report how your feature selection strategy was able to detect that it is irrelevant
- Add a new feature as the sum of two existing features to your extracted features and report how your feature selection strategy was able to detect that this new feature is dependent