In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Features and Feature Selection

Hamid R. Rabiee
Jafar Muhammadi

Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

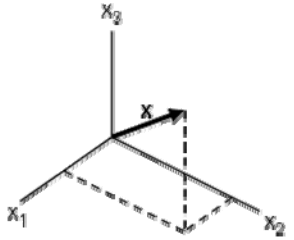
- **Features and Patterns**
  - **The Curse of Size and Dimensionality**
    - Features and Patterns
- **Data Reduction**
  - **Sampling**
  - **Dimensionality Reduction**
- **Feature Selection**
  - **Feature Selection Methods**
    - Univariate Feature selection
    - Multivariate Feature selection
  - **Right Method Picking**

Features and Patterns

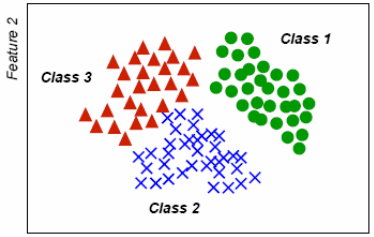
- **Feature**
  - **Feature is any distinctive aspect, quality or characteristic of an object (population)**
    - Features may be symbolic (e.g. color) or numeric (e.g. height).
- **Definitions**
  - **Feature vector**
    - The combination of d features is presented as a d-dimensional column vector called a feature vector.
  - **Feature space**
    - The d-dimensional space defined by the feature vector is called the feature space.
  - **Scatter plot**
    - Objects are represented as points in feature space. This representation is called a scatter plot.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature vector



Feature Space (3D)



Scatter plot (2D)

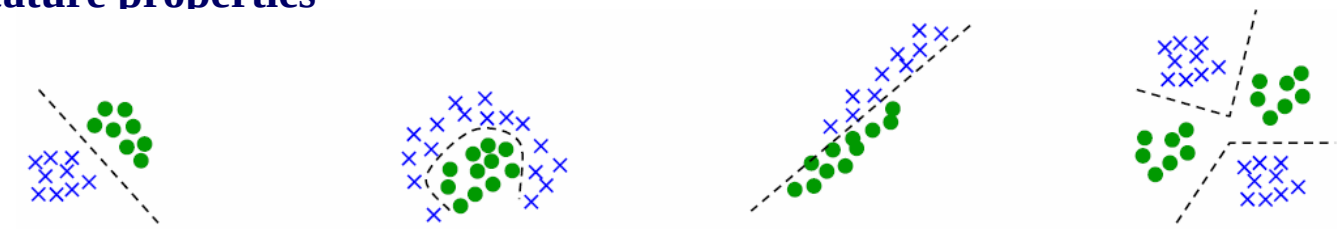
Features and Patterns

- **Pattern is a composite of traits or features corresponding to characteristics of an object or population**
  - **In classification; a pattern is a pair of feature vector and label**
- **What makes a good feature vector**
  - **The quality of a feature vector is related to its ability to discriminate samples from different classes**
    - Samples from the same class should have similar feature values
    - Samples from different classes have different feature values

Good features **Bad features**

Features and Patterns

- **More feature properties**

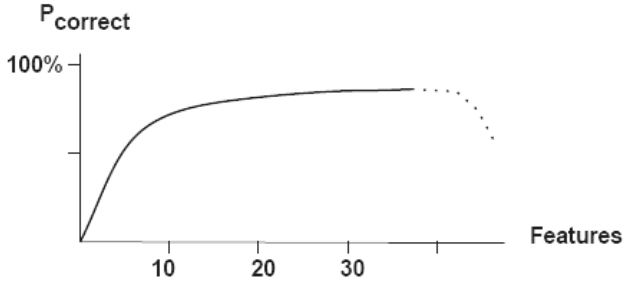


Linear Separability | Non-Linear Separability | Highly correlated | Multi modal

- **Good features are:**
  - Representative: provide a concise description
  - Characteristic: different values for different classes, and almost identical values for very similar objects
  - Interpretable: easily translate into object characteristics used by human experts
  - Suitable: natural choice for the task at hand
  - Independent: dependent features are redundant

The Curse of Size and Dimensionality

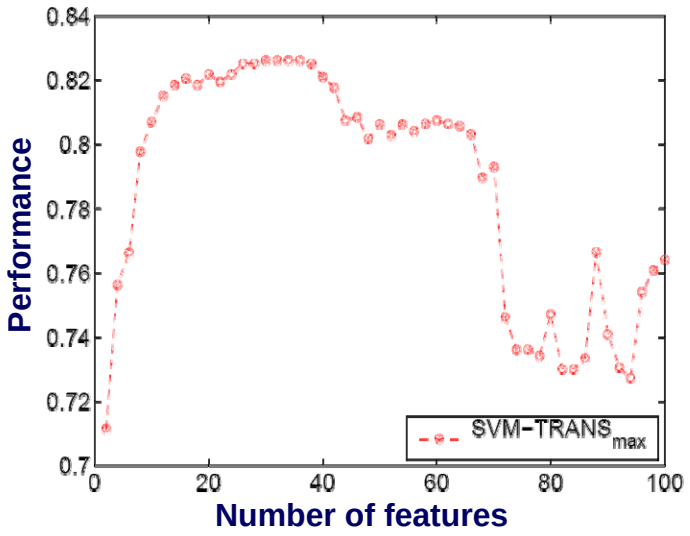
- **The performance of a classifier depends on the interrelationship between**
  - **sample sizes**
  - **number of features**
  - **classifier complexity**
- **The probability of misclassification of a decision rule does not increase beyond a certain dimension for the feature space as the number of features increases.**
  - **This is true as long as the class-conditional densities are completely known.**

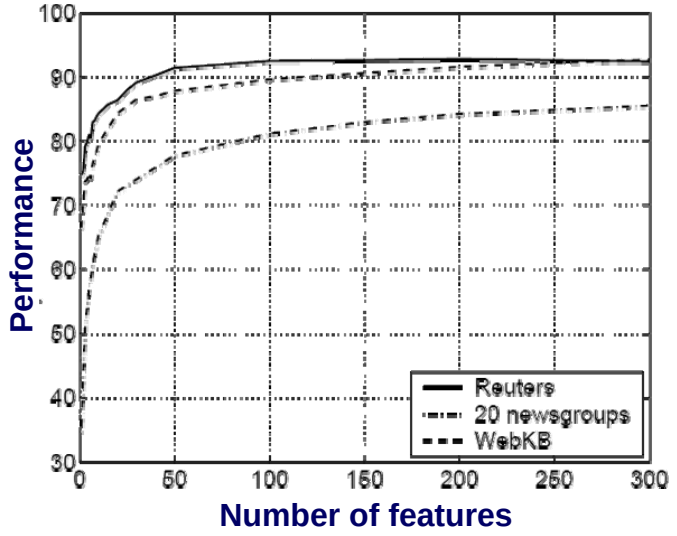


- **Peaking Phenomena**
  - **Adding features may actually degrade the performance of a classifier**

Features and Patterns

- **The curse of dimensionality examples**
  - **Case 1 (left): Drug Screening (Weston et al, Bioinformatics, 2002)**
  - **Case 2 (right): Text Filtering (Bekkerman et al, JMLR, 2003)**





Features and Patterns

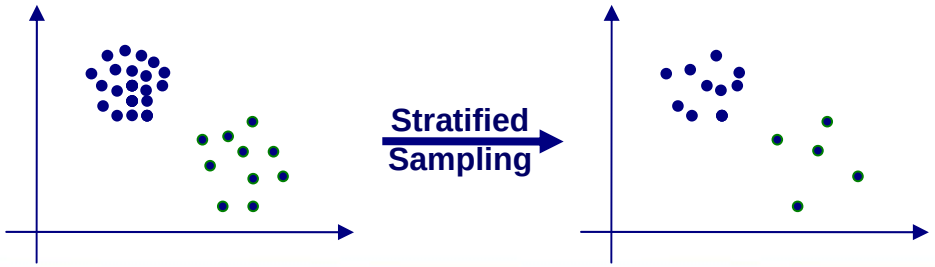
- **Examples of number of samples and features**
  - **Face recognition application**
    - For 1024*768 images, the number of features will be 786432 !
  - **Bio-informatics applications (gene and micro array data)**
    - Few samples (about 100) with high dimension (6000 – 60000)
  - **Text categorization application**
    - In a 50000 words vocabulary language, each document is represented by a 50000-dimensional vector
- **How to resolve the problem of huge data**
  - **Data reduction**
  - **Dimensional reduction (Selection or extraction)**

Data Reduction

- **Data reduction goal**
  - **Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results**
- **Data reduction methods**
  - **Regression**
    - Data are modeled to fit a determined model (e.g. line or AR)
  - **Sufficient Statistics**
    - A function of the data that maintains all the statistical information of the original population
  - **Histograms**
    - Divide data into buckets and store average (sum) for each bucket
    - Partitioning rules: equal-width, equal-frequency, equal-variance, etc.
  - **Clustering**
    - Partition data set into clusters based on similarity, and store cluster representation only
    - Clustering methods will be discussed later.
  - **Sampling**
    - obtaining small samples to represent the whole data set D

Sampling

- **Sampling strategies**
  - **Simple Random Sampling**
    - There is an equal probability of selecting any particular item
  - **Sampling without replacement**
    - As each item is selected, it is removed from the population, the same object can not be picked up more than once
  - **Sampling with replacement**
    - Objects are not removed from the population as they are selected for the sample.
      - **In sampling with replacement, the same object can be picked up more than once**
  - **Stratified sampling**
    - Grouping (split) population samples into relatively homogeneous subgroups
    - Then, draw random samples from each partition according to its size



Dimensionality Reduction

- **A limited yet salient feature set simplifies both pattern representation and classifier design.**
  - **Pattern representation is easy for 2D and 3D features.**
  - **How to make pattern with high dimensional features viewable? (refer to HW 1)**
- **Dimensionality Reduction**
  - **Feature Selection (will be discussed today)**
    - Select the best subset from a given feature set
  - **Feature Extraction (e.g. PCA etc., will be discussed next time)**
    - Create new features based on the original feature set
    - Transforms are usually involved

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \rightarrow \begin{bmatrix} x_{i_1} \\ \vdots \\ x_{i_m} \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \rightarrow \begin{bmatrix} y_{i_1} \\ \vdots \\ y_{i_m} \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}\right)$$

Feature Selection

✧ **Problem definition**

$m \leq d$, usually

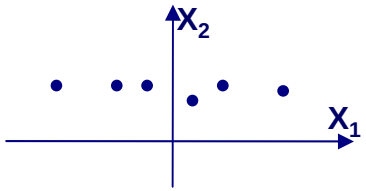
Number of possible Selections: $\binom{d}{m}$

Feature Selection Methods

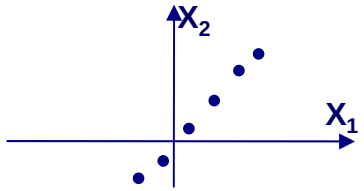
- **One view**
  - **Univariate method**
    - Considers one variable (feature) at a time
  - **Multivariate method**
    - Considers subsets of variables (features) together.
- **Another view**
  - **Filter method**
    - Ranks features subsets independently of the classifier.
  - **Wrapper method**
    - Uses a classifier to assess features subsets.
  - **Embedded**
    - Feature selection is part of the training procedure of a classifier (e.g. decision trees)

Univariate Feature selection

- **Filter methods have been used often (Why?)**
- **Criterion of Feature Selection**
  - **Criteria = Significant difference * Independence**
    - Significant difference: Pattern separability on individual candidate features
    - Independence: Non-correlation between candidate feature and already-selected features



X_1 is more significant than X_2



X_1 and X_2 both are significant, but correlated

- **Significant difference and independence can be used separately, too.**

Univariate Feature selection

- **Information based significant difference**
  - **select A if Gain(A) > Gain(B).**
  - **Gain can be calculated using several methods such as:**
    - Information gain, Gain ratio, Gini index (These methods will be discussed in TA session).
- **Statistical significant difference**
  - **Continuous data with normal distribution**
    - Two classes: T-test ← will be discussed here!
    - Multi classes: ANOVA
  - **Continuous data with non-normal distribution or rank data**
    - Two classes: Mann-Whitney test
    - Multi classes: Kruskal-Wallis test
  - **Categorical data**
    - Chi-square test

Univariate Feature selection

- **Independence**
  - **Based on correlation between a feature and a class label.**
    - $\text{Independence}^2 = 1 - \text{Correlation}^2$
  - **if a feature is heavily dependent on another, then it is redundant.**
  - **How calculate correlation?**
    - Continuous data with normal distribution
      - **Pearson correlation ← Will be discussed here!**
    - Continuous data with non-normal distribution or rank data
      - **Spearman rank correlation**
    - Categorical data
      - **Pearson contingency coefficient**

Univariate Feature selection

- **Univariate T-test**
  - **Select those features for which the means are significantly different for at least one pair of classes ➔ Using two sample t-test, and verifying the rejection of sameness hypothesis**
  - **Let x_{ij} be feature i of class j, with m_{ij} and s_{ij} as estimates of the mean and standard deviation, respectively. Use the statistic**

$$t = \frac{m_{ij} - m_{ik}}{\sqrt{s_{ij}^2 / N_j + s_{ik}^2 / N_k}}$$

where N_i is the number of feature vectors in class i. The statistic t will have a Student's t distribution, with degrees of freedom df, where

$$df = \frac{\left[s_{ij}^2 / N_j + s_{ik}^2 / N_k \right]^2}{\left(s_{ij}^2 / N_j \right)^2 / N_j + \left(s_{ik}^2 / N_k \right)^2 / N_k} - 2$$

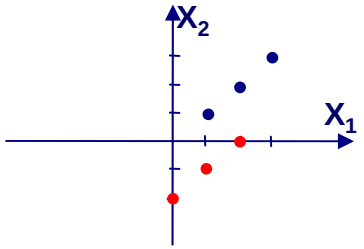
Hint: use closest integer, or interpolate.

  - **If $t < t_{\frac{\alpha}{2}}^{df}$ or $t > t_{1-\frac{\alpha}{2}}^{df}$ then reject the hypothesis that $m_{ij}=m_{ik}$.**

Univariate Feature selection

- **Univariate T-test example**
  - **Is X_1 is a good feature?**

$$m_{11} = 2,\ m_{12} = 1$$
$$s_{11}^2 = 1,\ s_{12}^2 = 1$$
$$\Rightarrow t = 1.224,\ df = 4 \Rightarrow t_{0.975}^4 = 2.78 > 1.224$$



Then, X_1 is not a good feature.

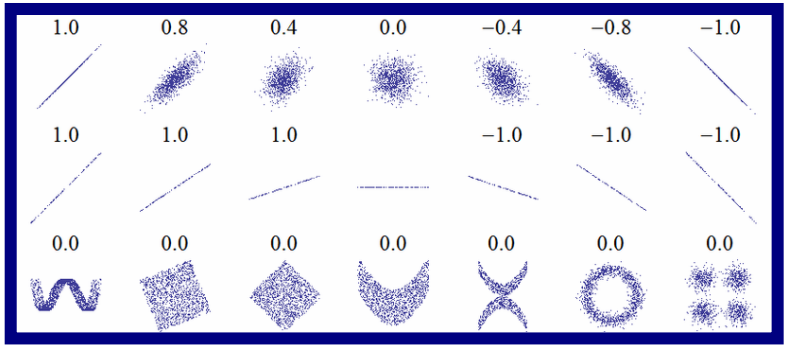
  - **Is X_2 is a good feature?**

$$m_{21} = 2,\ m_{22} = -1$$
$$s_{21}^2 = 1,\ s_{22}^2 = 1$$
$$\Rightarrow t = 3.674,\ df = 4 \Rightarrow t_{0.975}^4 = 2.78 < 3.674$$

Then, X_2 is a good feature.

Univariate Feature selection

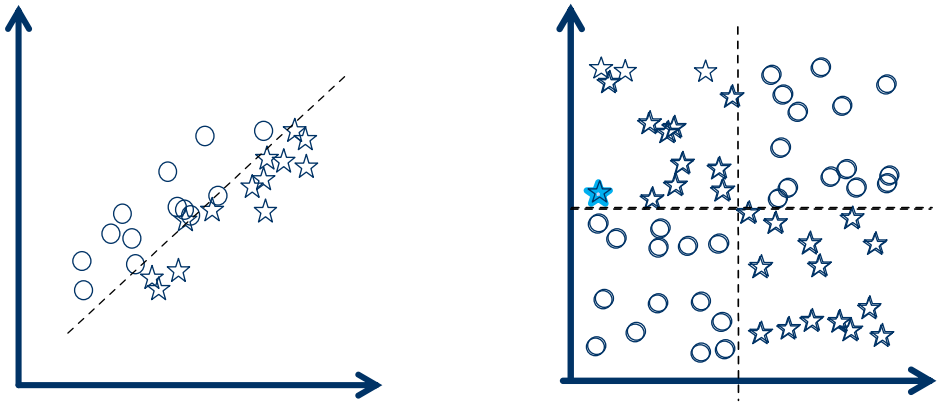
- **Pearson correlation**
  - **The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short).**
  - **The correlation between two variables reflects the degree to which the variables are related.**
  - **How to use Pearson correlation in order to decide that a feature is good one or not? Some heuristics are:**
    - Compute Pearson correlation between this feature and current selected ones
      - **Choose this feature if there isn't any high correlated feature in the selected ones.**
    - Compute the Pearson correlation between sample values and their output values (their classes)
      - **High correlation is preferred here (Why?)**
    - There are many papers of using correlation in order to feature selection.



$$\rho_{X,Y} = corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Univariate Feature selection

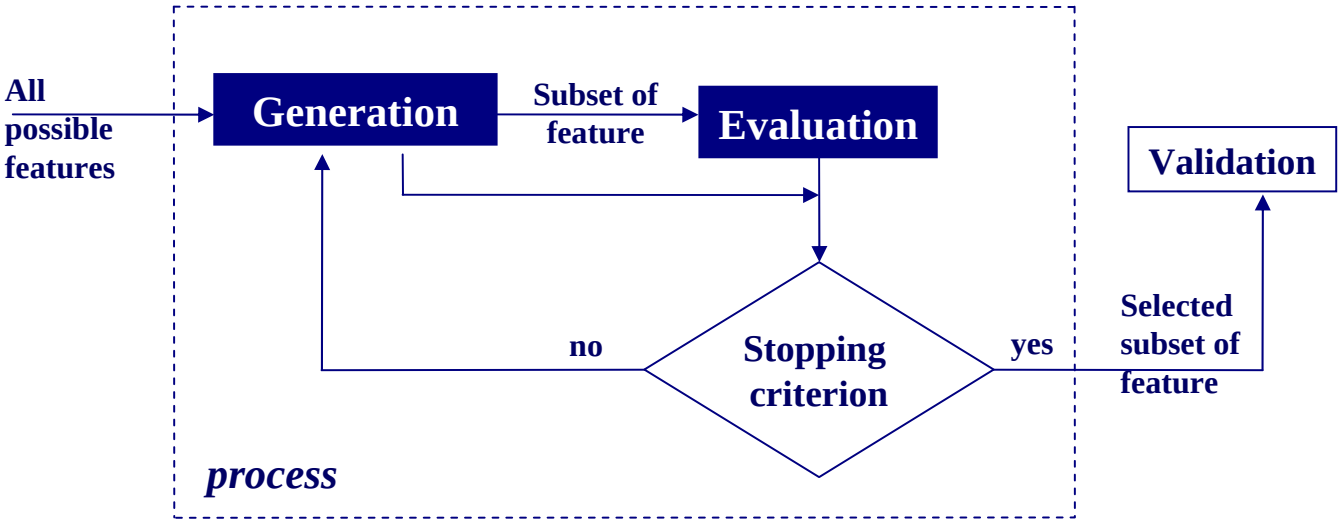
- **Univariate selection may fail!**
  - **Consider the following two example datasets:**



Guyon-Elisseeff, JMLR 2004; Springer 2006

Multivariate Feature selection

- **Multivariate feature selection steps**
  - **In the next slides we'll introduce common "Generation" and "Evaluations" methods.**



Multivariate Feature selection

- **Generation Methods**
  - **Complete/exhaustive**
    - Examine all combinations of feature subset (Too expensive if feature space is large).
    - Optimal subset is achievable.
  - **Heuristic**
    - Selection is directed under certain guideline
    - Uses incremental generation of subsets, often.
    - Possibility of miss out high importance features.
  - **Random**
    - no pre-defined way to select feature candidate. pick feature at random.
    - optimal subset depend on the number of tries
    - require more user-defined input parameters.
      - **result optimality will depend on how these parameters are defined.**

Multivariate Feature selection

- **Evaluation Methods**
  - **Filter methods**
    - Distance (Euclidean, Mahalanobis and etc.)
      - **select those features that support instances of the same class to stay within the same proximity.**
      - **instances of same class should be closer in terms of distance than those from different class.**
    - Consistency (min-features bias)
      - **Selects features that guarantee no inconsistency in data.**
      - **two instances are inconsistent if they have matching feature values but group under different class labels.**
      - **prefers smallest subset with consistency.**
    - Information measure (entropy, information gain, etc.)
      - **entropy - measurement of information content.**

Multivariate Feature selection

- **Evaluation Methods**
  - **Filter methods**
    - Dependency (correlation coefficient)
      - **correlation between a feature and a class label.**
      - **how close is the feature related to the outcome of the class label?**
      - **dependence between features is equal to degree of redundancy.**
  - **Wrapper method**
    - Classifier error rate (CER)
      - **evaluation function = classifier (loss generality)**

Multivariate Feature selection

- **Evaluation methods comparison criteria**
  - **Generality: how general is the method towards diff. classifiers?**
  - **Time: how complex in terms of time?**
  - **Accuracy: how accurate is the resulting classification task?**

Evaluation Methods	Generality	Time	Accuracy
Distance	Yes	Low	-
Information	Yes	Low	-
Dependency	Yes	Low	-
Consistency	Yes	Moderate	-
classifier error rate	No	High	Very High

Multivariate Feature selection

- **Algorithm design using introduced "generation" and "evaluate" methods**
  - **Relief Algorithm**
    - Generation: heuristic - evaluation: distance
  - **Branch & Bound Algorithm**
    - Generation: complete, evaluation: distance
  - **LVF Algorithm**
    - Generation: Random, evaluation: Consistency
  - **Decision Tree Method (DTM)**
    - Generation: Heuristic, Evaluation: Information
  - **Minimum Description Length Method (MDLM)**
    - Generation: Complete, Evaluation: Information
- **For more algorithms and details of them, refer to:**
  - **M. Dash, H. Liu, "Feature selection methods for classification", Intelligent Data Analysis: An International Journal, Elsevier, Vol. 1, No. 3, pp 131 – 156,1997.**

Any Question?

End of Lecture 3

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/