Statistical Pattern Recognition

Features and Feature Selection

Hamid R. Rabiee
Jafar Muhammadi

Spring 2014
http://ce.sharif.edu/courses/92-93/2/ce725-2/
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
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.
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 should have different feature values
Features and Patterns

✧ More feature properties

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

✧ Good features are (Consider the Fish example):

- Representative: provide a concise description (Like weight, Length)
- Characteristic: different values for different classes, and almost identical values for very similar objects (Like length and weight)
- Interpretable: easily translate into object characteristics used by human experts (Interpretable: Length, Weight; Uninterpretable: Length+Weight)
- Independent: dependent features are redundant (Independent: Length and color; dependent: Length and Weight)
The Curse of Size and Dimensionality

✧ The performance of a classifier depends on the interrelationship between
  ✧ sample sizes
  ✧ number of features
  ✧ classifier complexity

✧ Consider a 3-class pattern recognition problem
  ✧ 3 type of objects have to be classified
  ✧ A simple procedure:
    ✧ Divide the feature space in to uniform bins
    ✧ Compute the ratio of each class at each bin, and
    ✧ For a new example, find its bin and choose the predominant class in that bin
  ✧ We start with one feature and divide the real line into 3 bins
The Curse of Size and Dimensionality

✧ Moving to two dimensions.
  ✧ Increase the number of bins from 3 to \(3^2 = 9\)

✧ Moving to 3 features
  ✧ The number of bins grows to \(3^3 = 27\)
  ✧ For the same number of examples the 3D scatter plot is almost empty!
The Curse of Size and Dimensionality

✧ **Curse of dimensionality:**
  ✧ The number of bins grows exponentially as the Number of features increase.
  ✧ The exponential grows of number of sample data needed to ensure that the cells are not empty.

✧ **In practice the curse of dimensionality means that:**
  ✧ For a given sample size, there is a maximum number of features above which the performance of the classifier will degrade rather than improve
  ✧ Peaking Phenomena
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
    ✧ Feature Selection
    ✧ Feature 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 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 cannot 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

\[
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_d \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
    x_i \\
    x_{i+1} \\
    \vdots \\
    x_{i+m} \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_d \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
    y_1 \\
    y_{i+m} \\
\end{bmatrix}
= f(\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_d \\
\end{bmatrix})
\]
Feature Selection

✧ Problem definition

\[ X_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \]

\[ X'_i = (x_{i1}', x_{i2}', \ldots, x_{im}')^T \]

\[ m \leq d, \text{ usually} \]

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
  
  ✧ Significant difference

  ❀ Independence: Non-correlated feature selection

  ❀ Discrimination Power

\[ X_1 \text{ is more significant than } X_2 \]

\[ X_1, X_2 \text{ both are significant, but correlated} \]

\[ X_1 \text{ results in more discrimination than } X_2 \]
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.
    ✧ Independence² = 1 – Correlation²
  ✧ 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 significantly different features, one by one:
    - Difference in means → Hypothesis testing
  - Let $x_{ij}$ be feature $i$ of class $j$, with $m_{ij}$ and $s_{ij}$ as estimates of the mean and standard deviation, respectively. $N_i$ is the number of feature vectors in class $i$.

\[
\begin{align*}
H_0: \quad & m_{ij} = m_{ik} \\
H_1: \quad & m_{ij} \neq m_{ik}
\end{align*}
\]

\[
t = \frac{m_{ij} - m_{ik}}{\sqrt{s_{ij}^2/N_j + s_{ik}^2/N_k}} \sim \text{T-student}(df)
\]

Hint: use closest integer, or interpolate.

- If $t < t_{\alpha/2}^d$ or $t > t_{1-\alpha/2}^d$ then reject $H_0$. 

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

✧ Univariate T-test example

✧ Is $X_1$ is a good feature?

\[
\begin{align*}
    m_{11} &= 2, \quad m_{12} = 1 \\
    s_{11}^2 &= 1, \quad s_{12}^2 = 1 \\
    \Rightarrow t &= 1.224, \quad df = 4 \Rightarrow t_{0.975}^4 = 2.78 > 1.224
\end{align*}
\]

Then, $X_1$ is not a good feature.

✧ Is $X_2$ is a good feature?

\[
\begin{align*}
    m_{21} &= 2, \quad m_{22} = -1 \\
    s_{21}^2 &= 1, \quad s_{22}^2 = 1 \\
    \Rightarrow t &= 3.674, \quad df = 4 \Rightarrow t_{0.975}^4 = 2.78 < 3.674
\end{align*}
\]

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).

✧ **How to use Pearson correlation in order to decide that a feature is good one or not?**
  
  ✧ 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.

\[ \rho_{XY} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \]
Univariate Feature Selection

✧ Univariate selection may fail!

✧ Consider the following two example datasets:

Guyon-Elisseff, JMLR 2004; Springer 2006
Multivariate Feature Selection

✧ Multivariate feature selection steps

✧ In the next slides we’ll introduce common “Generation” and “Evaluations” methods.

Process:

All possible features → Generation → Subset of feature → Evaluation → Stopping criterion

Validation

Selected subset of feature
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?

<table>
<thead>
<tr>
<th>Evaluation Methods</th>
<th>Generality</th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>Yes</td>
<td>Low</td>
<td>-</td>
</tr>
<tr>
<td>Information</td>
<td>Yes</td>
<td>Low</td>
<td>-</td>
</tr>
<tr>
<td>Dependency</td>
<td>Yes</td>
<td>Low</td>
<td>-</td>
</tr>
<tr>
<td>Consistency</td>
<td>Yes</td>
<td>Moderate</td>
<td>-</td>
</tr>
<tr>
<td>classifier error rate</td>
<td>No</td>
<td>High</td>
<td>Very High</td>
</tr>
</tbody>
</table>

✧ For more algorithms and details of them, refer to:

Any Question?

End of Lecture 3

Thank you!

Spring 2014

http://ce.sharif.edu/courses/92-93/2/ce725-2/