Statistical Pattern Recognition: A Review

Anil K. Jain, Fellow, IEEE, Robert Duin, and Jianchang Mao, Senior member, IEEE

Presented By:

Ashish Kumar(Y5123)
Ashutosh Kumar(Y5128)

Objective

- To summarize and compare well known methods at various stages in statistical Pattern recognition.
- Identify applications at the forefront of this field.

Pattern Recognition(PR)

- How machines observe the environment
- How they learn to distinguish patterns of interest from the background
- How they make reasonable decisions about the categories
- What is a pattern?
  - Defined as “opposite of chaos” by Watanabe
  - Entity vaguely defined that could be given a name
  - e.g. fingerprint image, human face, speech signal

Examples of applications

- **Optical Character Recognition (OCR)**
  - Handwritten: sorting letters by postal code, input device for PDA's.
  - Printed texts: reading machines for blind people, digitalization of text documents.
- **Biometrics**
  - Face recognition, verification, retrieval.
  - Finger prints recognition.
  - Speech recognition.
- **Diagnostic systems**
  - Medical diagnosis: X-Ray, EKG analysis.
  - Machine diagnostics, waster detection.
- **Military applications**
  - Automated Target Recognition (ATR).
  - Image segmentation and analysis (recognition from aerial or satelite photographs).

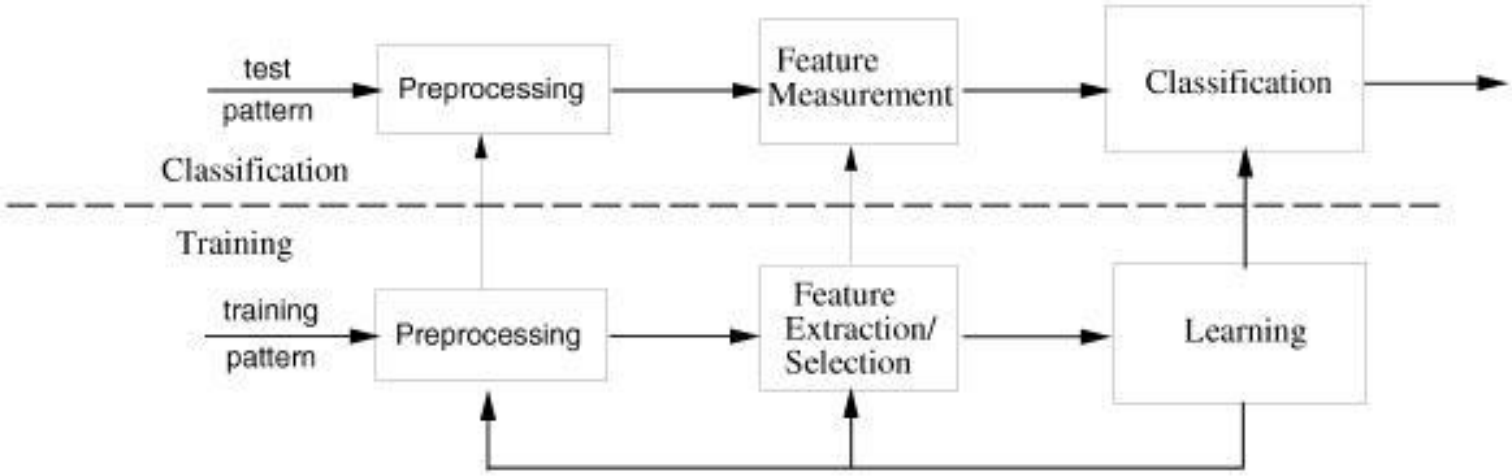
Approaches for PR

- Template Matching
  - Patterns are matched to stored templates
- Statistical Approach(SPR)
  - based on underlying statistical model of patterns and pattern classes.
- Syntactic Approach
  - pattern classes represented by means of formal structures as grammars, automata, strings, etc.
- Neural Networks
  - classifier is represented as a network of cells modeling neurons of the human brain (connectionist approach).

Statistical Pattern Recognition(SPR)

- Pattern represented by set of *d* features
- Decision boundaries formed using statistical decision theory
- Operated in two modes:
  - Training(learning)
  - Classification (testing)
- Major steps:
  - Preprocessing
  - Representation
  - Decision making

SPR(continued)



Decision Making

- Assignment of pattern to one of c classes w_1 , w_2 , …. , w_c .
- Vector of d features $\mathbf{x} = x_1 , x_2 , ….. , x_d$
- Features are assumed to have probability density/mass functions conditioned on pattern class
- Well known techniques:
  - Bayes Decision rule
  - Maximum likelihood rule (special case of Bayes rule)
  - Neyman-Pearson rule

Bayesian Decision Rule

- It involves minimizing the risk function.

$$R(\omega_i|\boldsymbol{x}) = \sum_{j=1}^{c} L(\omega_i,\omega_j) \cdot P(\omega_j|\boldsymbol{x})$$

- Where,
  - L(w_i ,w_j) is the loss incurred in deciding w_i when true class is w_j .
  - P(w_j |$\boldsymbol{x}$) is the posterior probability.
  - R is the risk function.

Bayesian Decision Rule

- An example of typical loss function (0/1 loss function)
  - $L(w_i, w_j) = 0$ when $i = j$

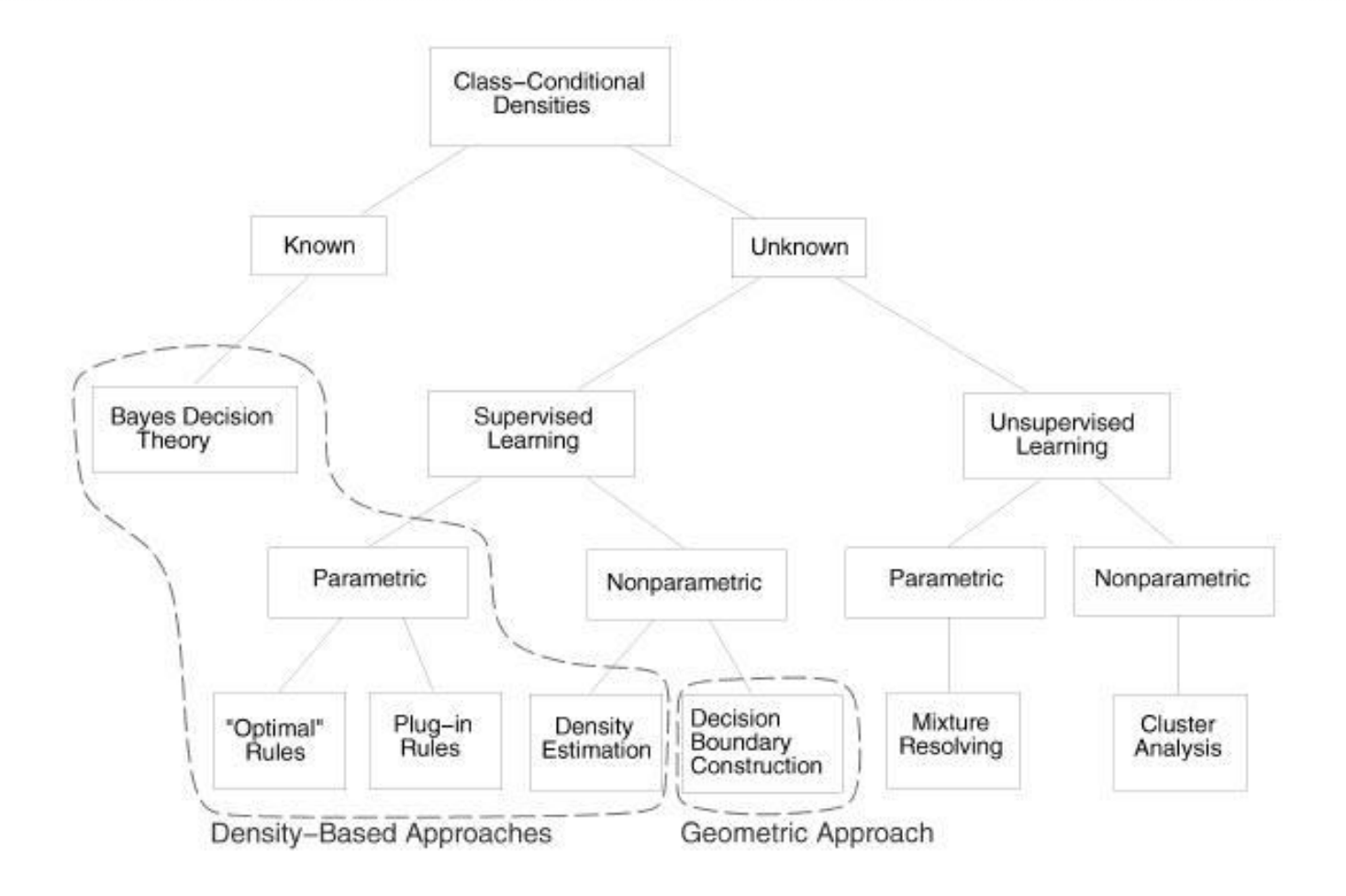
    $= 1$ *when* $i \neq j$
  - In this case, Bayesian decision rule can be simplified as follows (also called Maximum a posteriori(MAP) rule):

    x is assigned to class w_i if

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \text{ for all } j \neq i.$$

- The major limitation of the Bayesian rule:
  - Its demand for complete specification of class conditional probability densities.

Various Approaches in SPR

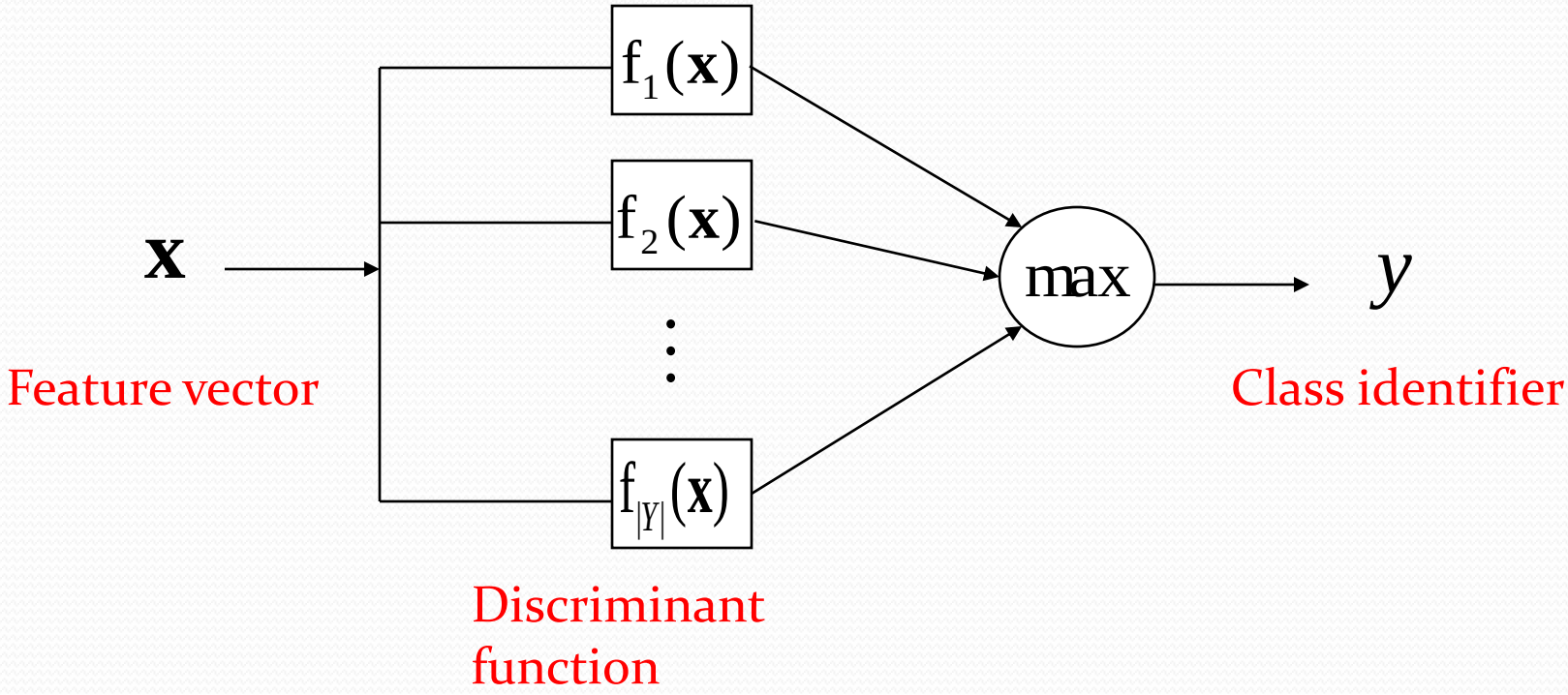


Representation of classifier

A classifier is typically represented as a set of discriminant functions

$$\mathrm{f}_i(\mathbf{x}) : X \to \Re, i = 1, \ldots, |Y|$$

The classifier assigns a feature vector **x** to the *i*-the class if $\mathrm{f}_i(\mathbf{x}) > \mathrm{f}_j(\mathbf{x}) \forall j \neq i$



Design of Classifiers

- Three Approaches:
  - Concept of Similarity
    - Need to establish a good metric to define similarity
    - Ex. Template Matching, minimum distance classifier
  - Probabilistic approach
    - Bayesian optimal/plug-in rules
    - Density estimation
  - Geometric approach
    - Directly creating decision boundaries by optimizing ceratin cost
    - Ex. Support Vector Machine(SVM)

Problems

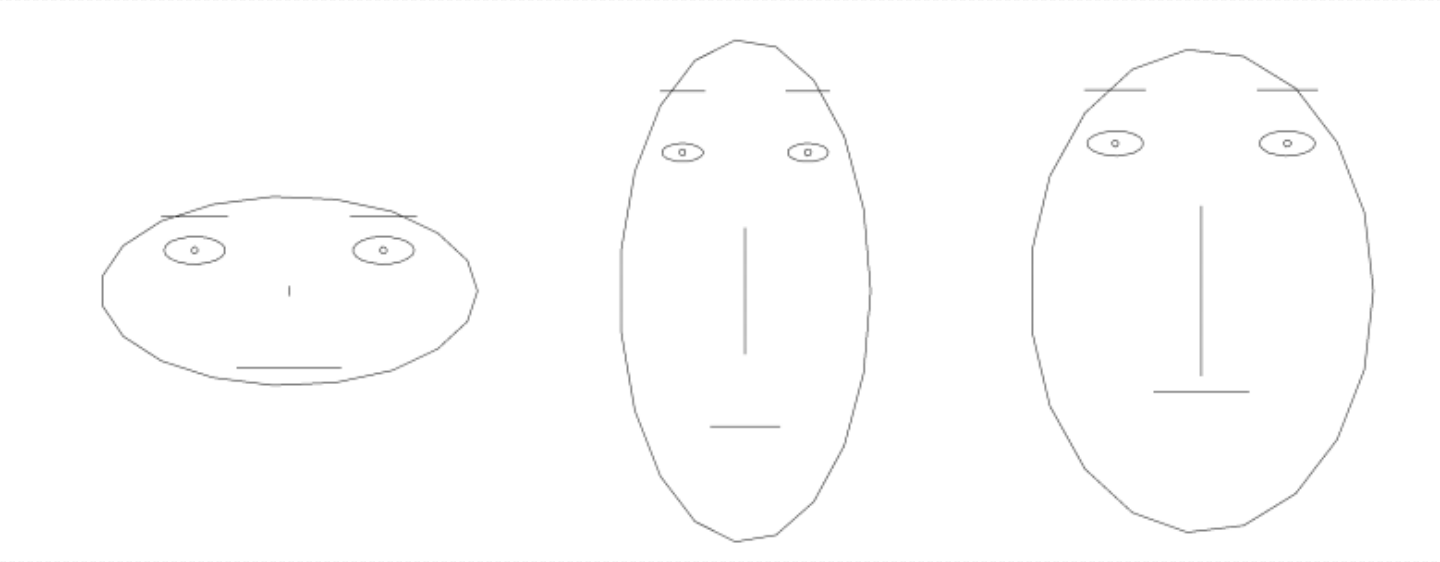
- The curse of Dimensionality
  - Added features reduce the efficiency of the classifier .
  - If number of training vectors are limited.
- parametric classifiers estimate the unknown parameters.
- For fixed sample size, as the number of features is increased ,number of unknown parameters increases .
- reliability of the parameter estimates decreases.

Dimensionality Reduction

- Problems in high dimensional feature vectors
  - Measurement cost
  - Curse of Dimensionality
  - Watanabe's "ugly duckling theoram" -two arbitrary patterns can be made similar with large number of redundant features.
- Reduce dimensionality
  - Feature Extraction
  - Feature Selection

Visual Examination

- Obtain a two- or three-dimensional projection of the given multi-dimensional data.
- eg. Representing each pattern as a cartoon face .



Feature Extraction

- Features are extracted from the sensed data.
- Linear transforms are used:
  - principal componentAnalysis(PCA),
  - factor analysis,
  - linear discriminant analysis.
- PCA
  - Orthogonal Projection of data onto a lower dimensional linear space.
  - Minimizes the mean squared between data points and their projections .

Feature Selection

- Feature selection leads to savings in measurement cost.
- X be a given set of features with cardinality of d.
- *Y be a subset of m features. [m<d]*
- Define J(Y) such that a higher value of J(Y) denotes a better feature subset.
  - Eg $J(Y) = (1-P_e)$
  - P_e is the classification error.

Unsupervised Pattern Recognition

- Construct decision boundaries based on unlabeled training data.
- Also known as data clustering.
- A cluster consists of a relatively high density of points separated from other clusters by a low density of points.
- Data Clustering :
  - Square-error partitional clustering.
  - agglomerative hierarchical clustering

Square-Error Clustering

- Minimizes the square-error

$$e_k^2 = \sum_{i=1}^{n_k} \left(x_i^{(k)} - m^{(k)} \right)^T \left(x_i^{(k)} - m^{(k)} \right).$$

$$E_K^2 = \sum_{k=1}^{K} e_k^2.$$

  - K is the number of clusters
  - m is the mean of cluster k.
  - X is the vector of feature vectors.
- Eg K-means algorithm.

Thank You