EECS 433 Statistical Pattern Recognition

Ying Wu

Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208

http://www.eecs.northwestern.edu/~yingwu

Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

What are Patterns?

- Random vectors
  - $x \in \mathbb{R}^n$
  - they are static
- Random processes
  - 1-D signals $s(t)$, e.g., time series
  - 2-D signals $I(u, v)$, e.g., images
  - 3-D signals $V(u, v, t)$, e.g., video
  - they are dynamic
- What is a pattern?
  - a prototype with variations
    - its polymorphism is reflected by its variations
    - the variation has a probabilistic distribution
  - for random vector data, it simply is its distribution
  - for random process data, it is very difficult! E.g.,
    - speech, music, ECG
    - image, microarray
    - video

How Do We Represent Patterns?

- Using templates and rules is far from enough
  - as a pattern is likely to exhibit large variations
- thus, a critical issue is to model its variations
  - i.e., learning from the data
  - this is clear for patterns of random vector data
  - and this is the center problem in classical statistical pattern recognition
  - parametric or non-parametric
  - Bayesian or non-Bayesian
- but this is still not quite clear for random process data
  - representing these data is quite challenging
  - features and descriptive models
  - generative models
  - pattern theory

Three Basic Problems in Statistical Pattern Recognition

Let's denote the data by $\mathbf{x}$. There are three basic problems in statistical pattern recognition:

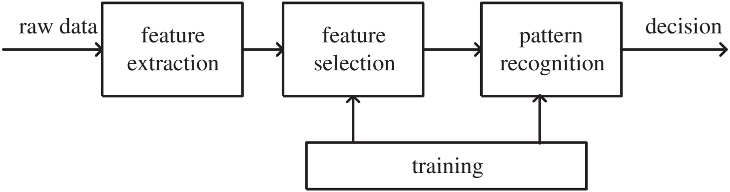
- Classification
  $f : \mathbf{x} \to C$, where C is a discrete set
- Regression
  $f : \mathbf{x} \to y$, where $y \in \mathbb{R}$ a continuous space
- Density estimation
  model $p(\mathbf{x})$ that is the probabilistic density of $\mathbf{x}$

Feature

- FEATURES can be regarded as the descriptors of the pattern
  - they are represented as vectors
  - in this sense, a pattern can be represented by a high dimensional random vector
  - curse of dimensionality
- features are extracted from raw data
  - weak features and strong features
  - domain knowledge
- features can also be learned
  - linear combination
    - optimal projection, e.g., PCA, LDA
    - sequential selection, e.g., boosting
  - nonlinear embedding
    - dimension reduction, e.g., ISOMAP, LLE, MDS
    - manifold learning, e.g., nonlinear PCA

Components in Pattern Recognition Systems

- Feature extraction
- Feature selection
- Recognizer
- Training



Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

Applications

- Speech recognition
- Optical character recognition
- Image-based object recognition
- Image-based biometrics
  - Face recognition/identification
  - Gender/age recognition
  - Gait recognition
  - Fingerprint recognition
  - Iris recognition
- Expression and emotion recognition
- Gesture and action recognition
- Video-based activity recognition
- Video-based event recognition and abnormality detection
- ...

Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

Supervised and Unsupervised Schemes

To train a classifier from data, there are two typical training schemes:

- Supervised learning
  - the training data has supervision, i.e., input-output pair
  - given a set of $\{\mathbf{x}_k, c_k\}_{k=1}^{N}$, to train the mapping $f : \mathbf{x} \to C$
  - the labeling acts as the teacher
  - e.g., LDA, SVM
- Unsupervised learning
  - no teacher, only a set of data $\{\mathbf{x}_k\}_{k=1}^{N}$
  - needs to find the clustering or grouping of the data
  - e.g., most clustering methods
  - it is difficult to determine the number of clusters for training
  - minimum description length (MDL) principle

Other Schemes

Besides supervised and unsupervised learning schemes, there are some others:

- Semi-supervised learning
  - a mixture of labeled and unlabeled data, $\{\mathbf{x}_k\}_{k=1}^{N_1} \bigcup \{\mathbf{x}_k, c_k\}_{k=1}^{N_2}$
  - only *critical* and *informative* data need to be labeled
  - what are those?
  - e.g., active learning
- Reinforcement learning
  - it has a teacher. More precisely, it is a critic
  - but the critic only tells *right* or *wrong*, but not how
  - such binary feedbacks are used to improve learning

Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

Bayesian and Non-Bayesian Approaches

- Bayesian approach
  - models *likelihood* $p(\mathbf{x}|C)$ and *prior* $p(C)$ distributions
  - computes *posterior* $p(C|\mathbf{x})$
  - based on Bayesian decision theory and *Bayes risk*

$$R(\alpha_k|\mathbf{x}) = \sum_{j=1}^{C} \lambda(\alpha_k|c_j)p(c_j|\mathbf{x})$$

  - leads to minimum-error-rate classification
  - Bayesian learning (ML estimation and Bayesian estimation)
- Non-Bayesian approach
  - empirical risk and overfitting
  - minimizing empirical risk $\rightarrow$ maximizing margin
  - minimizing training error $\rightarrow$ maximizing the generalizability
  - a great success of support vector machines

Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

Descriptive, Discriminative and Generative Models

- Descriptive models
  - features are derived from data
  - features are descriptive
  - it gives the likelihood of the observed data
  - e.g., PCA
- Discriminative models
  - features are derived from data
  - but they are *discriminative* to differentiate from other classes
  - it concentrates on the differences from others
  - e.g., LDA, SVM
- Generative models
  - explicitly models the formation or *generation* of the data
  - based on which the likelihood model is formed
  - no features derived from data directly
  - it can be used to *synthesize* data. Neither descriptive nor discriminative model can
  - e.g., Gaussian mixture models, HMM, MRF

Outline

What is Pattern Recognition?

Applications

Supervised and Unsupervised Schemes

Bayesian and Non-Bayesian Approaches

Descriptive, Discriminative and Generative Models

This Course

This Course

In this course, we'll cover the following topics:

- W-1: Bayesian decision and Bayesian classification, PCA/LDA
- W-2: ICA, Nearest neighbor classifiers
- W-3: Nonparametric density estimation, and linear discriminative models
- W-4: SVM and Kernel machines
- W-5: Feature selection and boosting
- W-6: EM, spectral clustering, sparsity models
- W-7: Metric learning, Deep neural networks, Dimension reduction and embedding
- W-8/9: Generative models
  - Bayesian networks (BN) and dynamic Bayesian nets (DBN)
  - Markov random field (MRF)
  - Conditional random field (CRF)

This Course

- Textbook
  - *Pattern Classification*, by R. Duda, P. Hart and D. Stork, 2nd Ed., Wiley-Interscience, 2001
  - Selected recent research papers
  - Lecture notes
- No midterm exam.
- Final projects
  - small group projects ($|group| \leq 2$)
  - suggestion: find a smaller topic and write a good comprehensive survey
  - 30-min talk and 20-page report