Chapter 1 Introduction

What is Pattern?

A pattern is the opposite of a chaos ; it is an entity vaguely defined, that could be given a name.

Various kinds of patterns

- Visual patterns (视觉模式) such as eyes, nose,mouth, face, fingerprint, etc.
- Temporal patterns (时序模式) such as speech,audios, videos, data streams, etc.
- Logical patterns (逻辑模式) such as characters,strings, images, etc.

What is Recognition?

Identification of a pattern as a member of a category we already know, or we are familiar with.

识别是将模式鉴定为我们已知或者熟悉的类别的成员

Two Types

Classification: Categories are **known** and the task is to assign a proper class label for each pattern

Clustering: Categories are **unknown** and the task is to learn categories and group the patterns accordingly

What is Pattern Recognition?

Pattern recognition is the procedure of **processing and analyzing diverse information** (numerical, literal, logical) characterizing the objects or phenomenon, so as to **provide descriptions, identifications, classifications and interpretations** for them.

对表征事物或现象的各种形式的（数值的，文字的和逻辑关系的）信息进行处理和分析，从而对事物或现象进行描述、辨认、分类和解释的过程。

Three procedures

- Perceive: : Observe the environment (e.g. interact with the real world)
- Process: Learn to distinguish patterns of interest from their background.
- Prediction: make sound and reasonable decisions about the categories of the patterns

Applications

- Character Recognition
- Speech Recognition
- Fingerprint Recognition
- Signature Recognition
- Face Detection
- Text categorization

Basic Concepts

Model

Descriptions which are typically mathematical in form

e.g. image to matrix; sound waves to frequency vector

Sample

Representatives of the patterns we want to classify

Training Set

A set of samples used to train classifiers

Test Set

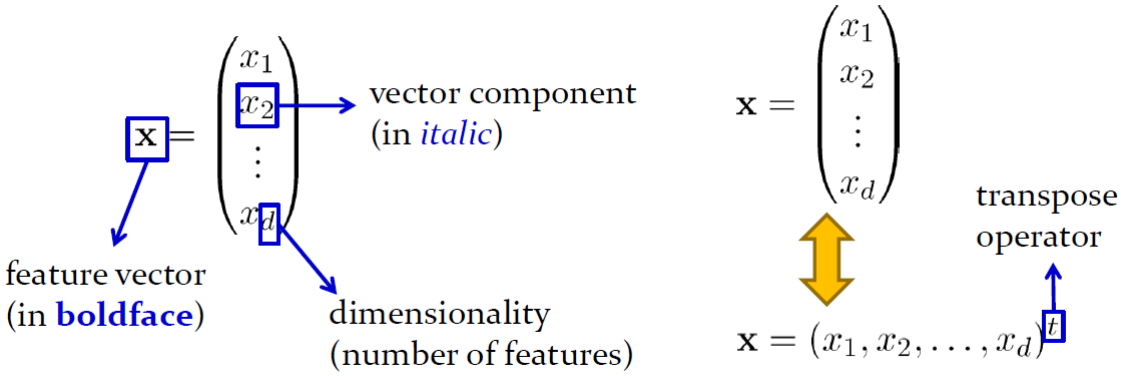
A set of samples to be classified, **usually being mutually exclusive to training set**

Feature

Attributes which characterize properties of the samples

Feature Vector

Vector formed by a group of features, **usually in column form**



Feature Space

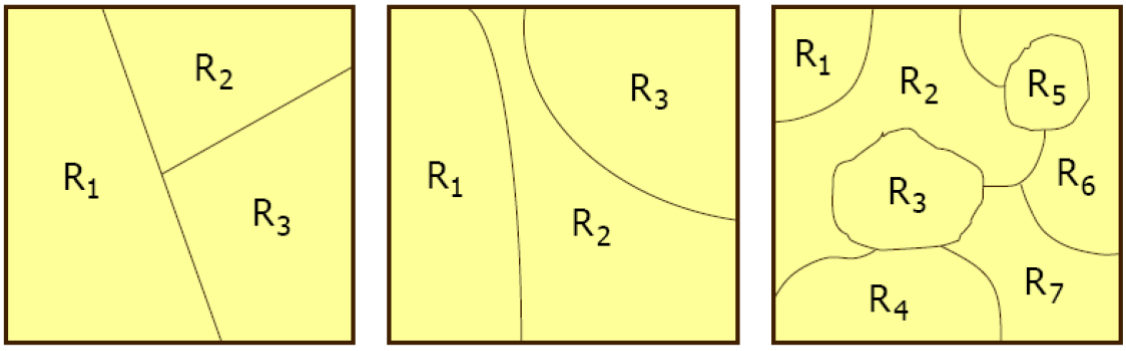
Space containing all the possible feature vectors

Scatter Plot

Each sample is plotted as a point in the feature space

Decision Boundary

Boundaries in feature space which separate different categories



linear boundary quadratic boundary complex boundary

Example: Sea Bass vs Salmon

Steps

- Preprocessing
- Feature Extraction
- Classification

Overfitting

Trade of between: Performance on training set / Simplicity of the classifier

Generalization (泛化能力)

The central aim of designing a classifier is to make correct decisions when presented with novel (unseen/test) patterns

The ULTIMATE goal

Related Fields to PR

- Pattern Recognition: Pattern to Category
- Hypothesis Testing: Null hypothesis to Rejection or Not
- Image Processing: Image to Image
- Regression: Pattern to Real Value
- Interpolation: Pattern (unexplored input range) to Interpolated Value
- Density Estimation: Patterns to Probability density function pdf for different categories

Pattern Recognition System

- Sensing: **converts** physical inputs into digital signal data
- Segmentation: **isolate** sensed objects from the background or from other objects
- Feature Extraction: **measures** object properties that are useful for classification
- Classification: **extracted** features to assign the sensed object to a category
- Post-processing: A post processor **decide on the appropriate action** based on the classification

Design Cycle of PR System

- Collect Data: A large part of the cost
- Choose Feature: highly domain-dependent / prior knowledge
- Choose Model
- Train Classifier
- Evaluate classifier

Important Issues

Noise

Definition

Any property of the sensed pattern which is not due to the true underlying model but instead to intrinsic randomness of the world or the sensors

- Various types of noise exist
- Noise can reduce the reliability of the feature values measured
- Knowledge of the noise process can help improve performance

Segmentation

Individual patterns have to be segmented for subsequent pattern recognition operations

One of the **deepest** and **hardest** problems

Different segmentations: e.g. BEATS $\rightarrow$ BE, BEAT, EAT, AT, EATS...

Data Collection

- A small set of "typical" examples $\rightarrow$ Preliminary study of system feasibility
- Much more data $\rightarrow$ Assure good performance in the fielded system

The Data collected: Is **adequately large**? Is **Representative**?

The efforts of data collection could be rather **demanding**

Domain Knowledge (prior knowledge)

Type I: Incorporate domain knowledge on the patterns themselves - HARD
To recognize all types of chairs — hard to find **commonness** for chairs

Type II: Incorporate domain knowledge on the pattern generation procedure
Optical character recognition (OCR) $\rightarrow$ Assume handwritten characters are written as a sequence of strokes
First try to recover stroke representations $\rightarrow$ deduce the character from the identified strokes

Feature Extraction

A domain dependent problem which influences the classifier's performance

Good extracted features $\rightarrow$ Make classification easier

Distinguishing Capability

Whose values are very **similar for objects in the same category**, while very **different for objects in different categories**

How to choose features?

- simple to extract
- robust to noise
- lead to simpler decision boundaries

Pattern Representation

Various ways for pattern representation:

- Statistical: feature vector (the most popular)
- Template Matching: prototype(原型) templates
- Syntactic(句法): rules or grammars

Desired Properties

- Patterns from the **same classes / different classes** should have **similar/different** representations
- Pattern representations should be **invariant to transformations** such as translations, rotations, resizes, reflections, non rigid deformations
- Intra-class/Inter-class variation should be small/large

Missing Features

In practical problems, values for certain features may be missing

Solutions

Naïve method: choose **zero / average** value

Sophisticated method: **regression** techniques

Model Selection

Each pattern recognition method employs certain **model hypothesis**

Every pattern recognition problem has its own **underlying true model** (not known)

Fundamental questions on model selection

- How do we know whether the hypothesized model **is (relatively) consistent with** the underlying true model?
- How are we to know to **reject a class** of models and try another one?
- Can we **automate the process of model selection**, instead of trial and error (试错) which is random and tedious?

Overfitting

We can get **perfect classification performance** on the training data by **choosing complex models**

Complex models are tuned to the **particular training samples** , rather than the **characteristics of the true model**

Context (上下文)

Input-dependent information , other than from the pattern itself

The same pattern within different context might have different meanings

Classifier Ensemble (分类器集成)

Classifier ensemble aims to **improve generalization performance** by **employing a number of classifiers** for the same task

Also known as Multi classifier System , Mixture of Experts , Classifier Fusion

Diverse ensemble techniques: Bagging, Boosting, Random subspace

Methods

- Majority voting: vote for the category where most classifiers agree
- Weighted voting: weight each vote by classifier's confidence
- Stacking: learn the rule of combination (more complicated)

Costs & Risks

Cost is the loss after making incorrect decisions

- Equal cost: In OCR, the cost of mistaking "6" as "9" might be equal to that of mistaking "9" as "6"
- Unequal cost: False Negative & False Positive

Risk is total expected cost which we want to optimize

- Error rate: percentages of test patterns being wrongly classified

Questions

- incorporate knowledge of costs
- estimate the lowest possible risk of any classifier

Computational Complexity

An algorithm scale with

- The number of features (dimensionality)
- The number of training patterns
- The number of possible categories

Brute force (蛮力) approaches might lead to perfect classification, but with **impractical time and storage requirements**