Pattern Recognition

Min Wang

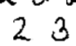
Textbook

- R. Duda, P. Hart, D. Stork, *Pattern Classification*, *second edition*, 2000（有中译本）
- 张学工，*模式识别（第三版）*，清华大学出版社，2010
- A.R.Webb, K.D.Copsey, Statistical Pattern Recognition, third edition, 2015（有中译本）

Application Cases

- Character Recognition
  - OCR (Optical Character Recognition)
- Speech Recognition
  - translation machine, identification
- Intelligent Traffic
  - License plate, car model
- Target Recognition
  - ATR (Automatic Target Recognition)
- Many more…

What is PR?

- Pattern 2
- Sample 
- Recognition 
- Learning/Training
- Evaluation/Test

5

What is PR?

- Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. (Anil K.Jain)

5

6

Visual data in life

6

7

Discovering visual patterns

7

8

Structured PR methods (Jingsun Fu, 1960s)

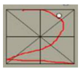
- Also called knowledge-based PR methods
- Structured representation (string, tree, graph)
- Structure (syntax) analysis
- Limited usage
- Difficulty in inference, recursion

8

9

Statistical PR methods (T.Pavlidis, 1971)

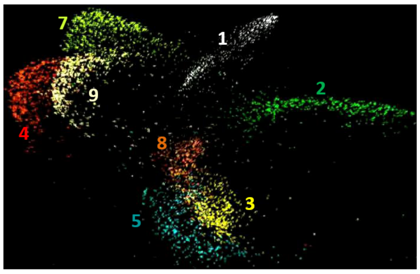
- A sample->a feature vector $x^T = [x_1,\ldots,x_n]$
- Feature vectors->feature space
- How to divide the feature space
- Widely used
- Less make use of the structure relationship of patterns



9

10

Handwritten digits recognition



Hand-written digit images
projected as points on a two-dimensional (nonlinear) feature spaces

10

11

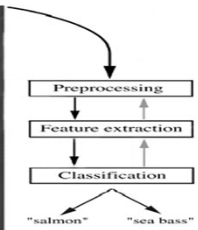
An Example

- "Sorting incoming fish on a conveyor according to species using optical sensing"

Species — Sea bass / Salmon

11

12



12

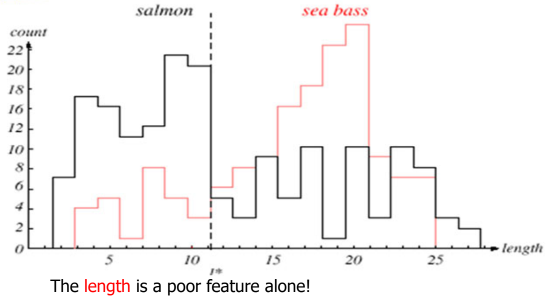
13

- Preprocessing
  - Use a segmentation operation to isolate fishes from one another and from the background
- Feature extraction
  - Reduce the Information from a single fish by measuring certain features
- Classification
  - Our focus

13

14

Classification by length

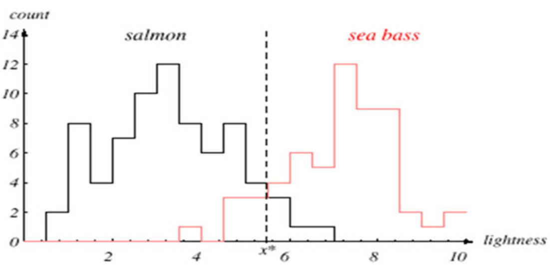


The length is a poor feature alone!
Select the lightness as a possible feature.

14

15

Classification by lightness



15

16

Decision theory

- Threshold decision boundary and cost relationship
  - Move our decision boundary toward smaller values of lightness in order to minimize the cost (reduce the number of sea bass that are classified salmon!)

Central task of decision theory

16

17

Feature vector

- Adopt the lightness and add the width of the fish

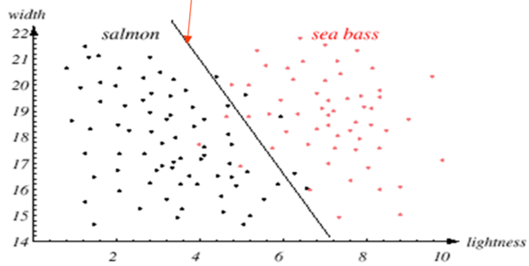
Fish ➡ $x^T = [x_1, x_2]$

Lightness Width

17

18

decision boundary



Underfitting!

18

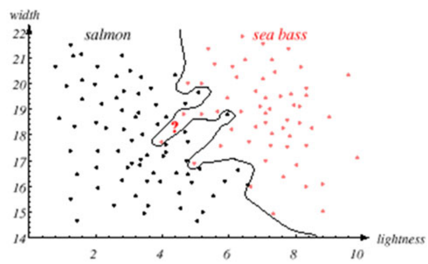
19

Feature choice

- We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such "noisy features"
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:

19

20



Overfitting!

20

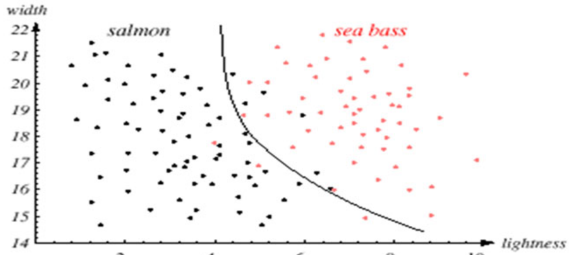
Generalization

- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

Issue of generalization!

21

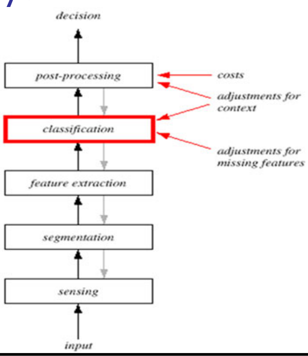
Model choice



Tradeoff!

22

A PR System



23

Sensing

- Use of a transducer (camera or microphone)
- PR system depends of the bandwidth, the resolution sensitivity distortion of the transducer
- Data collection. How do we know when we have collected an adequately large and representative set of examples for training and testing the system?
- Out of our scope

24

25

Segmentation

- Samples should be well separated and should not overlap
- Out of our scope

25

26

Feature extraction

- Feature choice
  - Discriminative features
  - Insensitive to noise
  - Invariant features with respect to translation, rotation and scale
  - Simple to extract
- Depends on the characteristics of the problem domain

26

27

Classification

- Use a feature vector provided by a feature extractor to assign the object to a category
- Model choice. Unsatisfied with the performance of our fish classifier and want to jump to another class of model
- Training. Use data to determine the classifier. Many different procedures for training classifiers and choosing models

27

28

Post-processing

- Measure the error rate (or performance) and switch from one set of features to another one
- Exploit context information other than from the target pattern itself to improve performance
- What is the trade-off between computational ease and performance?
- How an algorithm scales as a function of the number of features, patterns or categories?

28

29

Conclusion

- Reader seems to be overwhelmed by the number, complexity and magnitude of the sub-problems of Pattern Recognition
- Many of these sub-problems can indeed be solved
- Many fascinating unsolved problems still remain

29

30

Learning Algorithms

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Density Estimation
  - Clustering
  - Dimensionality Reduction

30

31

Supervised Learning

Feature Space $\mathcal{X}$ | **Label** Space $\mathcal{Y}$

Words in a document ⇨ "Sports" "News" "Science" … classification (Y discrete)

Market information up to time t ⇨ Share Price "$ 24.50" regression continuous)

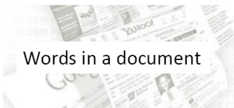
Given $D=\{\mathbf{X}_i, \mathbf{Y}_i\}$, learn $\mathrm{f}(\cdot): \mathbf{Y}_i = \mathrm{f}(\mathbf{X}_i)$, s.t. $D^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$

31

32

Unsupervised Learning- Density Estimation

Feature Space $\mathcal{X}$



Words in a document ⇨ Word distribution (Probability of a word)

Given $D=\{\mathbf{X}_i\}$, learn $\mathrm{f}(\cdot): \mathbf{Y}_i = \mathrm{f}(\mathbf{X}_i)$, s.t. $D^{\text{new}} = \{\mathbf{X}_j\} \Rightarrow \{\mathbf{Y}_j\}$

32

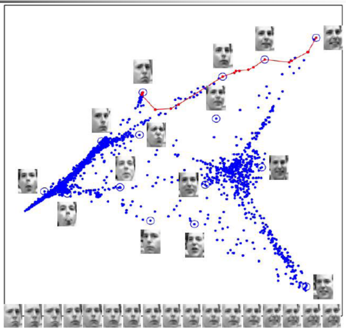
33

Clustering

33

34

Dimensionality Reduction



34

35

Supervised Learning vs. Unsupervised Learning

Supervised Learning – Learning with a teacher

$\{(X_i, Y_i)\}_{i=1}^n$ ⇨ Learning algorithm ⇨ $\widehat{f}_n$

Documents, topics

Mapping between Documents and topics

Unsupervised Learning – Learning without a teacher

$\{X_i\}_{i=1}^n$ ⇨ Learning algorithm ⇨ $\widehat{f}_n$

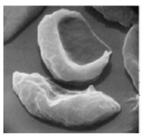
Documents

Model for word distribution OR Clustering of similar documents

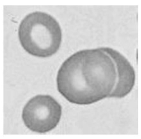
35

36

Performance Measures



⇨ "Anemic Cell" (0)

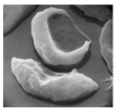


⇨ "Healthy Cell" (1)

36

37

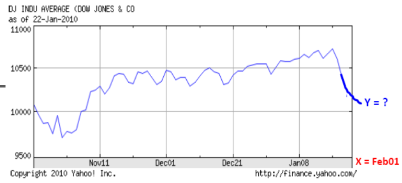
$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	*Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
	"Anemic cell"	"Anemic cell"	0
		"Healthy cell"	1

$$\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$$ **0/1 loss**

37

38



X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"$24.50"	"$24.50"	0
		"$26.00"	1?
		"$26.10"	2?

$$\text{loss}(Y, f(X)) = (f(X) - Y)^2$$ **square loss**

38

39

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

Don't just want label of one test data (cell image), but any cell image $X \in \mathcal{X}$

$$(X, Y) \sim P_{XY}$$

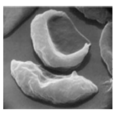
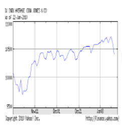
Given a cell image drawn randomly from the collection of all cell images, how well does the predictor perform on average?

$$\text{Risk } R(f) \equiv \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

39

40

Performance: Risk $R(f) \equiv \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$

		$\text{loss}(Y, f(X))$	Risk $R(f)$
	"Anemic cell"	$1_{\{f(X) \neq Y\}}$ **0/1 loss**	$P(f(X) \neq Y)$ **Probability of Error**
	Share Price "$ 24.50"	$(f(X) - Y)^2$ **square loss**	$\mathbb{E}[(f(X) - Y)^2]$ **Mean Square Error**

40

Bayes optimal rule

Ideal goal: Construct **prediction rule** $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^* = \arg\min_f \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

Bayes optimal rule

Best possible performance:

Bayes Risk $\quad R(f^*) \leq R(f) \quad$ for all f

BUT… Optimal rule is not computable - depends on unknown P_{XY} !

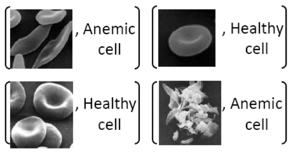
41

Experience—Training Data

Training data (experience) provides a glimpse of P_{XY}

(observed) $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$ **(unknown)**

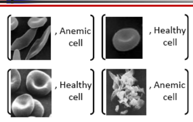
independent, identically distributed



Provided by expert, measuring device, some experiment, …

42

Supervised Learning Algorithm



$\Rightarrow$ **Learning algorithm** $\Rightarrow \widehat{f}_n$

Training data $\{(X_i, Y_i)\}_{i=1}^n$

$\widehat{f}_n$ is a mapping from $\mathcal{X} \to \mathcal{Y}$ $\quad \widehat{f}_n$  = "Anemic cell"

Test data X

Note: test data ≠ training data

43

What you should know…

- pattern,sample,recognition,
  pattern recognition
- feature,feature vector,
  feature space,decision boundary
- components of a PR system
- task,performance,experience
- Bayes optimal rule
- training (data),test (data),generalization

44