UNIVERSITY OF SURREY

UDRC Summer School

Pattern Recognition

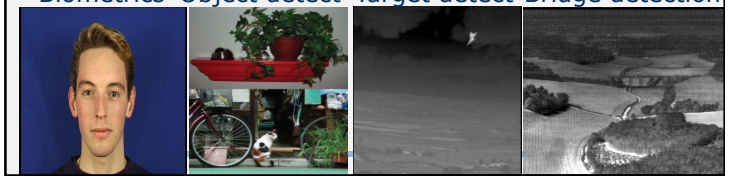
Josef Kittler
email: J.Kittler@surrey.ac.uk
www.ee.surrey.ac.uk/Personal/J.Kittler

UNIVERSITY OF SURREY

Introduction

- Pattern recognition is a branch of signal processing that focuses on the recognition of patterns and regularities in data
- Detected regularities (shape, texture, relations) enable signal classification
- Applications span a vast range of problems auto-mating human perception in decision making

Biometrics Object detect Target detect Bridge detection



UNIVERSITY OF SURREY

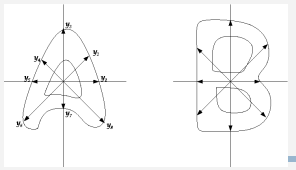
Pattern recognition problem variants

- Multiclass pattern recognition
- Detection (two class problem)
- One-class problems (anomaly detection)
- Verification
- Multilabel classification
- Retrieval

UNIVERSITY OF SURREY

Pattern recognition system

Z - input pattern
y - pattern representation vector
x - feature vector
d - decision



Pattern recognition problem

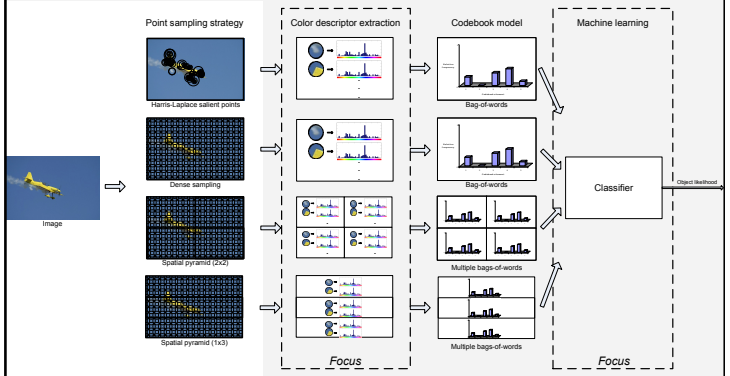
m – number of classes
x - pattern vector
$P(\omega_i)$ -priori probability of class ω_i
$p(\mathbf{x}|\omega_i)$ -measurement distribution of patterns in class ω_i

Sensor

- Broad concept, including
  - Physical device
  - Signal reconstruction and conditioning
  - Background removal
  - Registration
  - Invariant representation
- Sensor design/development is application specific

5

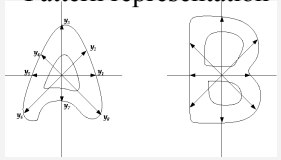
Example: Image analysis



6

Pattern recognition system

Pattern representation



Z - input pattern

y - pattern representation vector

x - feature vector

d - decision

Pattern recognition problem

m – number of classes

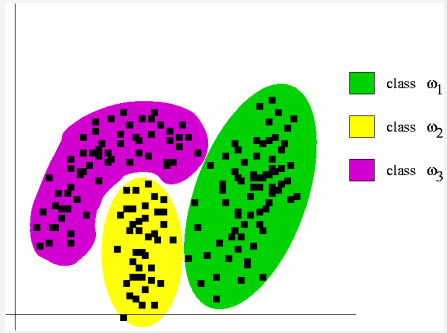
$P(\omega_i)$ -priori probability of class ω_i

$p(\mathbf{x}|\omega_i)$ -measurement distribution of patterns in class ω_i

7

Training data

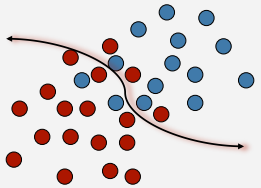
Geometric viewpoint of the pattern recognition problem



8

Overlapping classes

- Probabilistic model required



Course topics

- Statistical Pattern Recognition
  - Pattern generation process
  - Statistical decision theory
  - Density function estimation
  - Basic classifiers
  - Sparse representation
  - Similarity based classification
  - Performance evaluation
  - Multiple classifier systems
- Dimensionality reduction
  - Feature selection
  - Feature extraction
- Support Vector Machines (John Shawe-Taylor)
- Deep Neural Networks (Mark Plumbley)

UDRC Summer School

Statistical Pattern Recognition:

Classification

Basic probability relationships

- In Pattern Recognition we are dealing with two random variables
  class ω
  and pattern (measurement) vector $\mathbf{x}$
- The probability of their joint occurrence can be expressed in terms of conditional probabilities

Bayes formula relating conditional probabilities

$$p(\mathbf{x}, \omega) = p(\mathbf{x}|\omega)P(\omega) = P(\omega|\mathbf{x})p(\mathbf{x})$$

where $P(\omega|\mathbf{x})$ ..... aposteriori class probability
$p(\mathbf{x})$ ..... mixture density (measurement distribution), which can be computed as

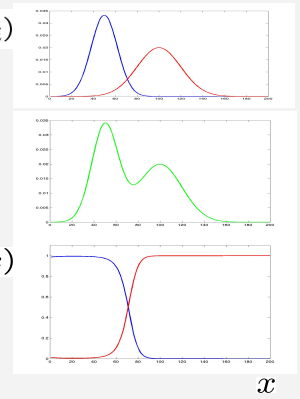
$$p(\mathbf{x}) = \sum_{\omega} p(\mathbf{x}|\omega)P(\omega)$$

Probability distribution functions

$p(\mathbf{x}|\omega_i)P(\omega_i)$

$p(x)$

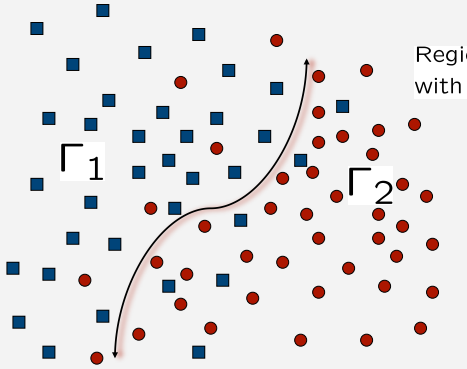
$P(\omega_i|x)$



x

13

Problem formulation



Region Γ_i associated with class ω_i

14

Statistical decision theory

- Given the probabilistic model of a pattern generating process, how do we decide the class membership of pattern $\mathbf{x}$
- Need to introduce decision costs ρ_{ij} associated with the assignment of pattern $\mathbf{x}$, belonging to class ω_i to class ω_j
- Note

ρ_{ii}	cost of correct decision
ρ_{ij}	cost of incorrect decision
$\rho_{ij} \geq 0$	
$\rho_{ij} \geq \rho_{ii}$	

- Example: *Signature verification*

ω_1	authentic signature
ω_2	forged signature
$\rho_{21} \geq \rho_{12}$	

15

Bayes minimum error rule

- For zero-one costs, i.e.

$$\rho_{ii} = 0 \quad \forall i$$
$$\rho_{ij} = 1 \quad \forall i,j \quad i \neq j$$

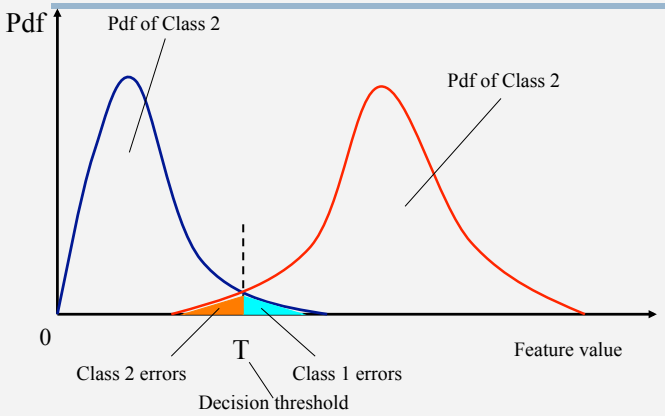
the r.h.s. of the Bayes minimum cost rule becomes:

$$\sum_{i=1}^{m} \rho_{ik} P(\omega_i|\mathbf{x}) = 1 - P(\omega_k|\mathbf{x})$$

and the corresponding decision rule is :

$$assign \quad \mathbf{x} \rightarrow \omega_j \quad if \quad P(\omega_j|\mathbf{x}) = \max_k P(\omega_k|\mathbf{x})$$

16



17

Actual decision rule

- Aposteriori class probabilities estimated from *training data set*
- Normally, the training data set would be labelled

$$X = \{\mathbf{x}_i, \gamma_i \; i = 1, \ldots N\}$$

- γ_i is a true class label provided by supervisor (teacher)
- System performance characterisation based on an independent *test data set*

$$X_T = \{\mathbf{x}_i, \beta_i \; i = 1, \ldots N\}$$

18

Terminology

- The design of a pattern recognition system using labelled data is referred to as *supervised learning*
- If class labels are unavailable, we deal with a *non supervised learning* problem
- PR system design involves
  - development of practical decision rules
  - model inference
  - performance evaluation

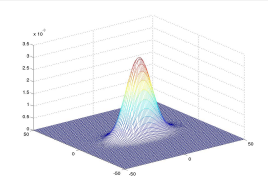
19

Parametric decision rules

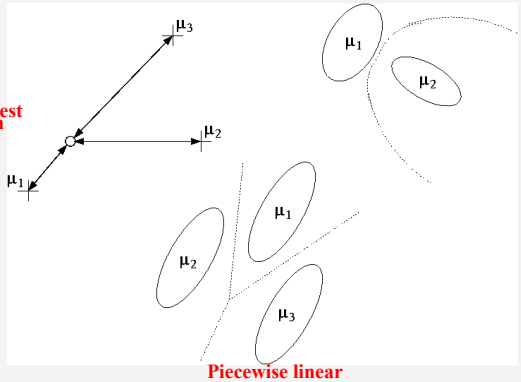
Gaussian classifiers

$$p(\mathrm{x}|\omega_i) = [(2\pi)^n|\Sigma_i|]^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathrm{x}-\mu_i)^T\Sigma_i^{-1}(\mathrm{x}-\mu_i)\}$$

- μ_i is the mean vector of class i
- Σ_i is the covariance matrix of class ω_i



Parametric decision rules



Nearest mean

Quadratic

Piecewise linear

k-nearest neighbour decision rules

Let $\mathbf{X} \equiv \{\mathbf{x}_j, \gamma_j | j = 1, ..., N\}$ be a training set of N labelled patterns

$$\mathbf{x} \rightarrow \omega_j \quad if \quad k_j = \max_{i=1}^{m} k_i$$

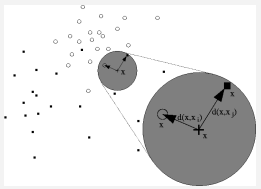
k_i….# of samples from class i among the k-nearest neighbours to **x**

- choice of k

nearest neighbour (k=1)

Properties of NN rule:

- intuitive
- suboptimal
- computationally expensive
- dependent on metric used



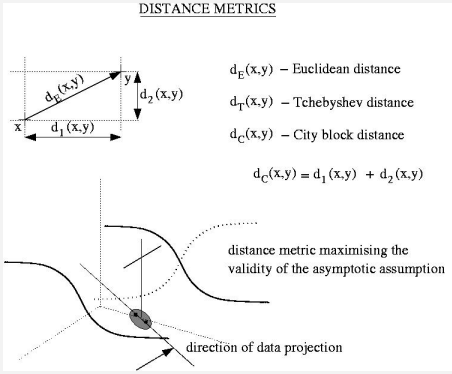
NN decision rule summary

$$assign \ \mathbf{x} \rightarrow \omega_j \ \ if \delta(\mathbf{x}_h, \mathbf{x}) = \min_{i}^{N} \delta(\mathbf{x}_i, \mathbf{x}) \ \ and \ \ \beta_h = \omega_j$$

where $\delta(x, z)$ is distance between two patterns and β_i is the label of pattern $\mathbf{x}_i$

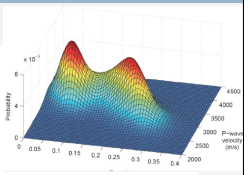
- Intuitively appealing
- Suboptimal performance unless the training set is edited using the MULTIEDIT algorithm
- Computationally involved, unless data set edited and condensed
- The choice of metric used for measuring distance is very important

Optimal metric



Gaussian Mixture models

- Let $\mathbf{x}_i, i = 1, ..., N$ be training set samples
- K.... a number of Gaussian components
- Estimate probability density as



$$p(\mathbf{x}) = \sum_{j=1}^{K} \pi_j f_j(\mathbf{x}|\theta_j) \quad (1)$$

- $f_j(\mathbf{x}|\theta_j)$ is a Gaussian component with parameters $\theta_j = (\mu_j, \Sigma_j)$ with μ_j and Σ_j denoting the mean and covariance matrix of the Gaussian component
- π_j is the j-th component weight

25

GMM estimation

- Given K, the parameters θ_j can be estimated using the Expectation Maximisation algorithm
- Initialise π_j, θ_j arbitrarily
- Compute

$$\begin{aligned} P(i|\mathbf{x}_t) &= \frac{\pi_i f_i(\mathbf{x}_t|\theta_i)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_t|j)} \\ n_i &= \sum_{t=1}^{N} P(i|\mathbf{x}_t) \\ \mu_i &= \frac{1}{n_i} \sum_{t=1}^{N} P(i|\mathbf{x}_t)\mathbf{x}_t \\ \Sigma_i &= \frac{1}{n_i} \sum_{t=1}^{N} P(i|\mathbf{x}_t)(\mathbf{x}_t - \mu_i)(\mathbf{x}_t - \mu_i)^T \\ \pi_i &= \frac{n_i}{N} \end{aligned} \quad (2)$$

26

GMM comments

- Local minima problem $\rightarrow$ different initialisations
- Choice of covariance matrix
- Choice of K
  - Optimise a joint criterion measuring fit to data subject to a model complexity penalty term (Bayesian Information Criterion)
  - Use predictive validation
  - Most recent approach based on the Dirichlet process model. No need to specify the number of components K

27

Dirichlet prior

- Based on the Chinese restaurant metaphor
- Allowing potentially an infinite number of components $K = \infty$ with a prior distribution over these components defined by a stick braking prior of the Dirichlet process

$$\pi_i = \nu_i \prod_{j=1}^{i-1}(1 - \nu_j) \quad (3)$$

where random variables ν_i are drawn from distribution $\beta(1, \alpha)$

- $\alpha \geq 1$ is a hyperparameter influencing the complexity of the fitted model, can be estimated from data
- $\pi_i, i = 1, ..., K$ specify the probabilities of the current components of the mixture and the probability of an additional component being created
- Choose a resonably large truncation parameter κ

28

Dirichlet prior GMM estimation

- Input: $X, \alpha, \beta(1,\alpha), \kappa$
- Compute $\pi_j, \forall j$ and initialise θ_j
  1. update model components

$$\begin{aligned} P(i|\mathbf{x}_t) &= \frac{\pi_i f_i(\mathbf{x}_t|\theta_i)}{\sum_{j=1}^{\kappa} \pi_j f_j(\mathbf{x}_t|j)} \\ n_i &= \sum_{t=1}^{N} P(i|\mathbf{x}_t) \\ \mu_i &= \frac{1}{n_i} \sum_{t=1}^{N} P(i|\mathbf{x}_t)\mathbf{x}_t \\ \Sigma_i &= \frac{1}{n_i} \sum_{t=1}^{N} P(i|\mathbf{x}_t)(\mathbf{x}_t - \mu_i)(\mathbf{x}_t - \mu_i)^T \end{aligned} \quad (4)$$

  2. update stick breaking probabilities

$$\begin{aligned} \nu_i &= \frac{n_i}{n_i + \alpha - 1 + \sum_{h=i+1}^{\kappa} n_h} \\ \pi_i &= \nu_i \prod_{j=1}^{i-1}(1-\nu_j) \end{aligned} \quad (5)$$

Sparse representation classification (SRC)

- Let $X = [\mathbf{x}_1, ...\mathbf{x}_N]$ be training data matrix
- Reconstruct an unknown sample $\mathbf{y}$ as

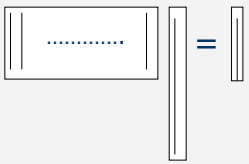
$$\mathbf{y} = X\mathbf{a} \quad (6)$$

  where $\mathbf{a}$ is a vector of coefficients
- If dimensionality $D < N$, a reconstruction solution could be found using LS, otherwise the solution has to be regularised by imposing a minimum norm on a
- Premise: for a sample $\mathbf{y}\epsilon\omega_i$, we would expect the reconstruction to be constituted by training samples from class i, i.e. vector $\mathbf{a}$ should be sparse (all entries for samples from other classes should be zero)

SRC

- By imposing sparsity on the reconstruction solution, we should be able to identify the class of $\mathbf{y}$
- This can be achieved by using l_1 and solving

$$argmin||\mathbf{a}||_1 \quad s.t. \mathbf{y} = X\mathbf{a} \quad (7)$$



SRC algorithm

1. Solve

$$argmin||\mathbf{a}||_1 \quad s.t. ||\mathbf{y} - X\mathbf{a}|| < \epsilon \quad (8)$$

2. Let $\mathbf{a_i}$ be vector $\mathbf{a}$ with all entries associated with samples from class $j \neq i$ set to zero, and compute the residual

$$r_i(\mathbf{y}) = ||\mathbf{y} - X\mathbf{a}_i||_2 \quad (9)$$

3. assign $\mathbf{y} \rightarrow \omega_i$ if $r_i(\mathbf{y}) = min_j r_j(\mathbf{y})$

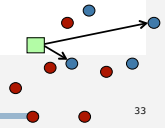
- Relationship to the k-NN classifier
  - k-NN classifier minimises the distances to $\mathbf{y}$
  - in addition, SRC involves pairwise interactions of residual error vectors

Similarity based classification

- NN classifier labels patterns based on similarity, gauged in terms of distance
- Squared Euclidean distance between **x** and **y** is given as

$$(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \qquad (10)$$

- The distance between **x** and **y** minimal when $\mathbf{x}^T\mathbf{y}$ is maximal
- scalar product $\mathbf{x}^T\mathbf{y}$ gauges similarity
- other notions of similarity can be defined, e.g. cosine similarity, Gaussian kernel, subjective grading



33

Similarity based class.

- Characterise objects in terms of their similarity to other objects, rather than by a set of measurements
- Let $o_i, i = 1, ..., N$ be an exemplar (training) set of objects and denote by $S(o_i, o_j)$ the similarity of objects o_i and o_j
- In the similarity based approach to pattern recognition, an unknown object o is characterised by vector **y** defined as

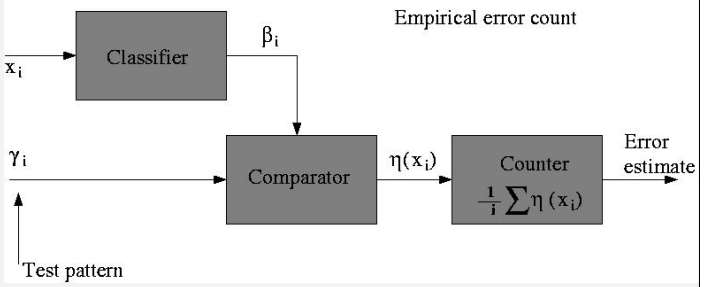
$$\mathbf{y} = [S(o, o_1), S(o, o_2), ..., S(o, o_N)]^T \qquad (11)$$

- The classification problem is solved in the feature space **y**
- The object measurement space is defined only implicitly
- There is a relationship with the kernel methods
- Similarity function $S(o, o_j)$ must be computable

34

Performance characterisation

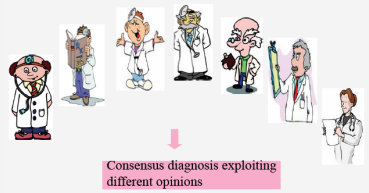
CLASSIFICATION ERROR ESTIMATION



35

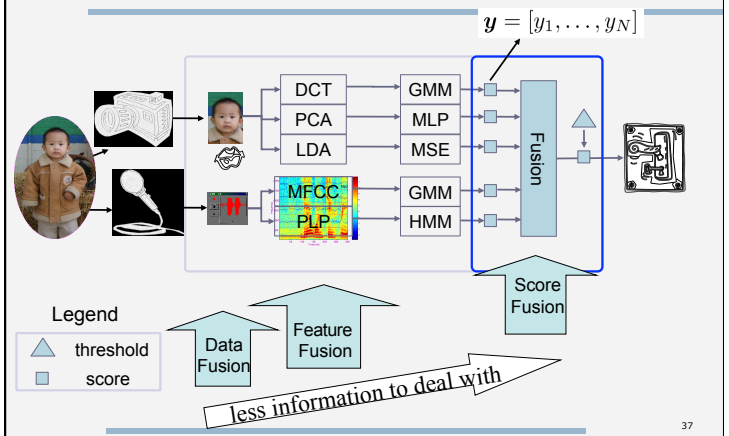
Multiple classifier systems

- Aim: fuse multiple classifier outputs to improve performance, combating the effect of
  - Particular training set
  - Particular classifier choice
  - Particular classifier parameters
  - Particular feature space
- Exploiting
  - Multiple modalities
  - Context
  - Quality gauging
  - Error correcting coding



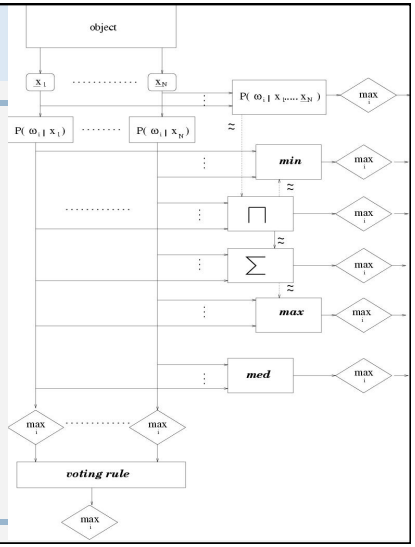
36

Levels of Fusion



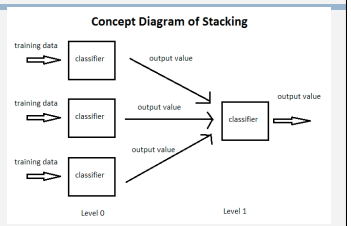
37

Fixed decision fusion strategies



Other examples

- Stacked generalisation (Wolpert)



- Error correcting output coding (ECOC) (Dietrich)

class

$$\begin{bmatrix} 100100110 \\ 011001001 \\ 101001010 \\ 100110001 \\ 001010011 \\ 110010100 \\ 001101001 \end{bmatrix}$$

code

39

References

1. PA Devijver and J Kittler, Pattern recognition: A statistical approach, Prentice Hall 1982
2. T Dietterich et al, Solving multiclass learning problems via error-correcting output coding, Jnl. Artificial Intel. Research, 1995
3. R Duda et al, Pattern classification, Wiley 2001
4. K Fukunaga, Introduction to statistical pattern recognition, Academic Press, 1990
5. T Kimura et al, Expectation-maximization algorithms for inference in Dirichlet processes mixture, Pattern Analysis and Appl. 16, pp 55-67, 2013
6. J Kittler et al, On combining classifiers, PAMI, pp 226-239, 1998
7. LI Kuncheva, Combining pattern classifiers, Wiley 2004
8. M Pelillo, Similarity-based pattern analysis and recognition, Springer 2013
9. C Rasmussen, The infinite Gaussian mixture model, In Advances in Neural Information Processing Systems 12, MIT Press 2000
10. AR Webb & KD Copsey, Statistical Pattern Recognition, Wiley 2011
11. DH Wolpert, Stacked generalization, Neural Networks,5, pp 241-259, 1992
12. AY Wright et al., Robust face recognition via sparse representation. IEEE PAMI, 2009.

40