The Dissimilarity Representation for Classification

INIT/AERFAI Summer School on MACHINE LEARNING
Benicassim, 25 June 2013

Robert P.W. Duin, Delft University of Technology, http://rduin.nl/

(In cooperation with Elżbieta Pękalska)

Pattern Recognition Lab
Delft University of Technology
The Netherlands

http://37steps.com/disrep-course/

Contents

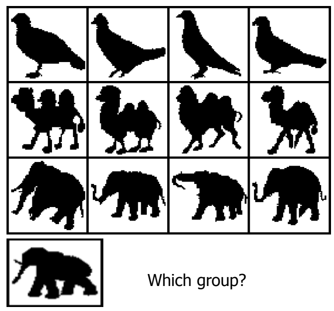
Pattern Recognition Problems

Blob Recognition

A

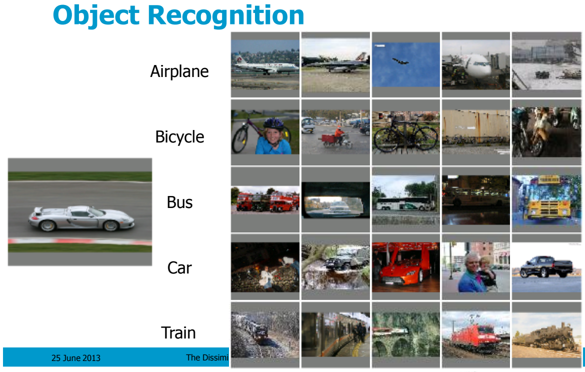
B

C

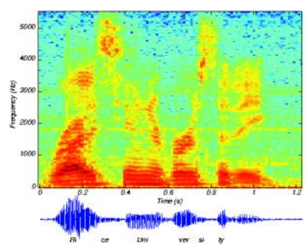


Which group?

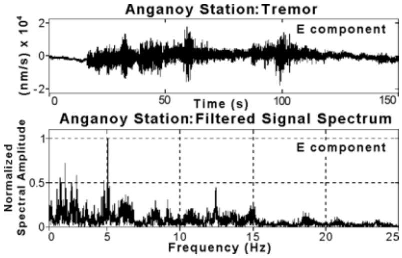
Object Recognition



Pattern Recognition: Speech

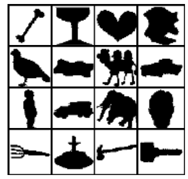


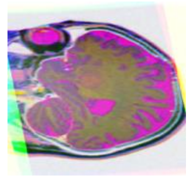
Pattern Recognition: Seismics



Earthquakes

Pattern Recognition Problems





To which class belongs an image

To which class (segment) belongs every pixel?

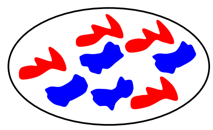
Where is an object of interest (detection); What is it (classification)?

Pattern Recognition: Shape Recognition

Pattern Recognition is very often Shape Recognition:

- Images: B/W, grey value, color, 2D, 3D, 4D
- Time Signals
- Spectra

Pattern Recognition: Shapes



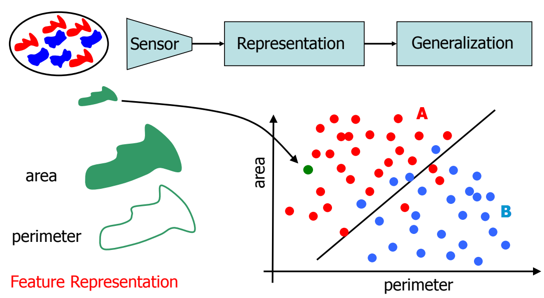
Examples of objects for different classes

Object of unknown class to be classified

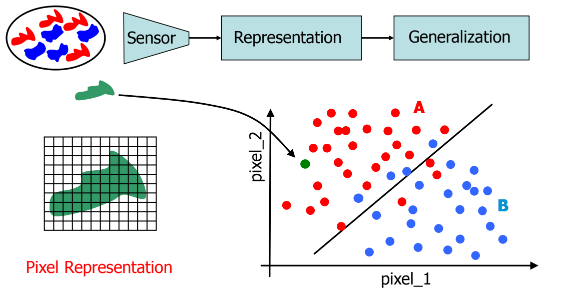
A ? B

Vector Representation

Pattern Recognition System



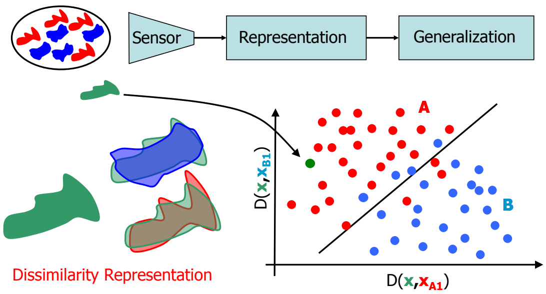
Pattern Recognition System



Pixel Representation

TUDelft

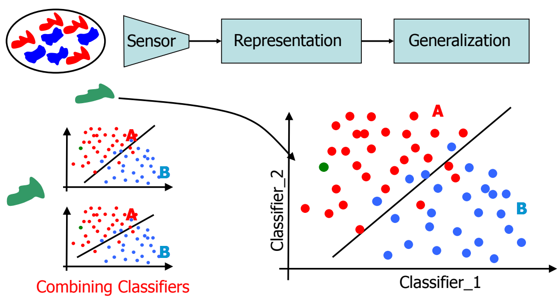
Pattern Recognition System



Dissimilarity Representation

TUDelft

Pattern Recognition System



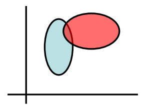
Combining Classifiers

TUDelft

Good Representations

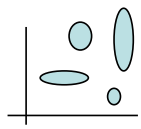
- Class specific

  Different classes should be represented in different positions in the representation space.



- Compact

  Every class should be represented in a small set of finite domains.

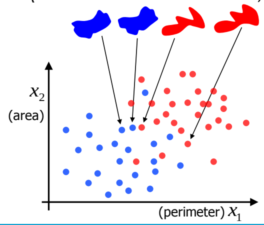


TUDelft

Compactness

Representations of real world similar objects are close.
There is no ground for any generalization (induction) on representations that do not obey this demand.

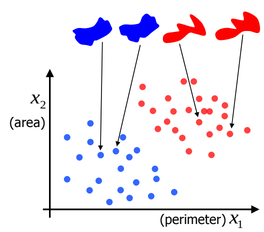
(A.G. Arkedev and E.M. Braverman, Computers and Pattern Recognition, 1966.)



The compactness hypothesis is not sufficient for perfect classification as dissimilar objects may be close.
→ class overlap
→ probabilities

TUDelft

True Representations



Similar objects are close
and
Dissimilar objects are distant.

→ no probabilities needed, domains are sufficient!

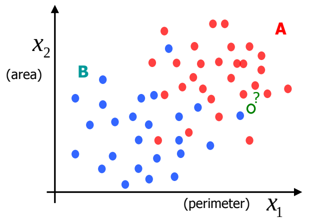
TUDelft

Distances and Densities

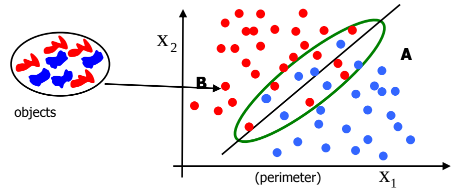
? to be classified as

B – because it is most close to an object B

A – because the local density of A is larger.



Features Reduce

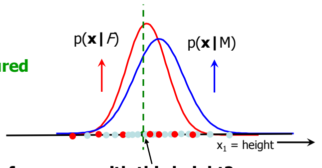


Due to reduction essentially different objects are represented identically.

→ The feature space representation needs a statistical, probabilistic generalization

Probabilistic Generalization

x = height measured



What is the gender of a person with this height?

Best guess is to choose the most 'probable' class (→ small error).

⇒ Good for overlapping classes.

⇒ Assumes the existence of a probabilistic class distribution and a representative set of examples.

Bayes decision rule, formal

$$p(A|x) > p(B|x) \rightarrow \text{A else B}$$

Bayes: $$\frac{p(x|A)\,p(A)}{p(x)} > \frac{p(x|B)\,p(B)}{p(x)} \rightarrow \text{A else B}$$

$$p(x|A)\,p(A) > p(x|B)\,p(B) \rightarrow \text{A else B}$$

2-class problems: $S(x) = p(x|A)\,p(A) - p(x|B)\,p(B) > 0 \rightarrow$ A else B

n-class problems: $\text{Class}(x) = \text{argmax}_{\omega}(p(x|\omega)\,p(\omega))$

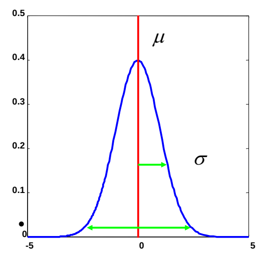
Density estimation

- The density is defined on the whole feature space.
- Around object x, the density is defined as:

$$p(x) = \frac{dP(x)}{dx} = \left(\frac{\text{fraction of objects}}{\text{volume}}\right)$$

- Given n measured objects, e.g. person's height (m) how can we estimate $p(x)$?

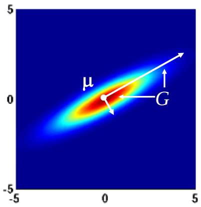
The Gaussian distribution (3)



- Normal distribution = Gaussian distribution
- Standard normal distribution: $\mu = 0,\ \sigma^2 = 1$
- 95% of data between $[\ \mu - 2\sigma,\ \mu + 2\sigma\ ]$ (in 1D!)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Multivariate Gaussians



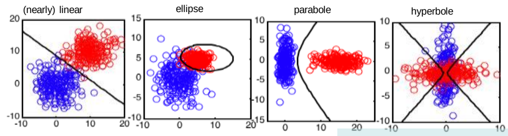
$$G = \begin{bmatrix} 3 & 1½ \\ 1½ & 2 \end{bmatrix}$$

- k - dimensional density:

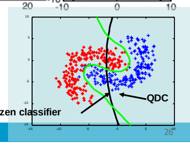
$$p(x) = \frac{1}{\sqrt{2\pi^k \det(G)}} \exp\left(-\frac{1}{2}(x-\mu)^T G^{-1}(x-\mu) \right)$$

Quadratic discriminant functions

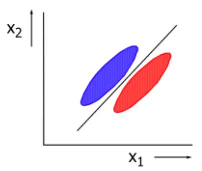
$$\boxed{R(x) = -\frac{1}{2}(x-\hat{\mu}_A)^T \hat{\Sigma}_A^{-1}(x-\hat{\mu}_A) + \frac{1}{2}(x-\hat{\mu}_B)^T \hat{\Sigma}_B^{-1}(x-\hat{\mu}_B) + \text{const}}$$



QDC assumes that classes are normally distributed. Wrong decision boundaries are estimated if this does not hold.

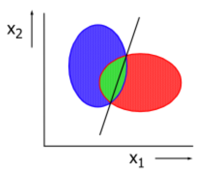


Linear discriminant function



Normal distributions with equal covariance matrices Σ are optimally separated by a linear classifier

$$R(x) = (\mu_A - \mu_B)^T \Sigma^{-1} x + \text{const}$$



Optimal classifier for normal distributions with unequal covariance matrices Σ_A and Σ_B can be approximated by:

$$R(x) = (\mu_A - \mu_B)^T (p(A)\Sigma_A + p(B)\Sigma_B)^{-1} x + \text{const}$$

Parzen density estimation (1)

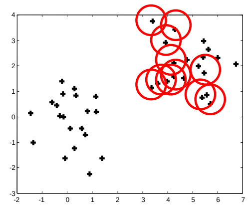
- Fix volume of bin, vary positions of bins, add contribution of each bin
- Define 'bin'-shape (kernel):

$$K(\mathbf{r}) > 0$$

$$\int K(\mathbf{r})\, d\mathbf{r} = 1$$

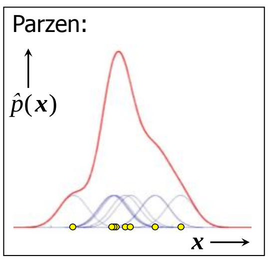
- For test object z sum all bins

$$p(\mathbf{z}) = \frac{1}{hn}\sum_i K\left(\frac{\mathbf{z} - \mathbf{x}_i}{h}\right)$$



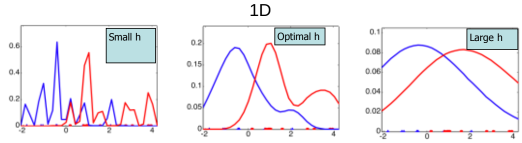
Parzen density estimation (2)

- With Gaussian kernel: $K(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right)$



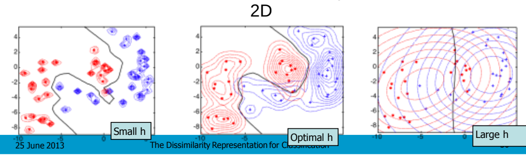
Parzen: density estimates vs the smoothing parameter

1D



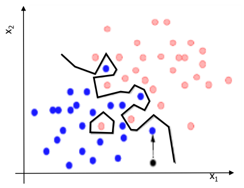
Increasing smoothing parameter h

2D



Nearest neighbor rule (1-NN rule)

Assign a new object to the class of the nearest neighbor in the training set.

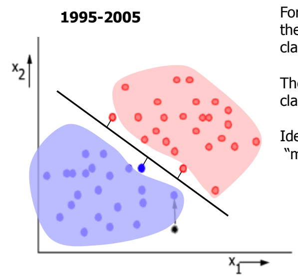


1-NN rule:

- Often relies on the Euclidean distance. Other distance measures can be used.
- Insensitive to prior probabilities!
- Scaling dependent. Features should be scaled properly.

There are no errors on the training set. The classifier is overtrained.

Support vector machine (SVM)



For linearly-separable classes find the few objects that determine the classifier. These are **support vectors.**

They have the same distance to the classifier: **the margin.**

Identical to "maximum-margin classifier"

$$S(x) = \sum_i \alpha_i (x_i^T x)$$

$$S(x) = w^T x, \min(w^T w)$$

Pixel Representation

Measuring Human Relevant Information

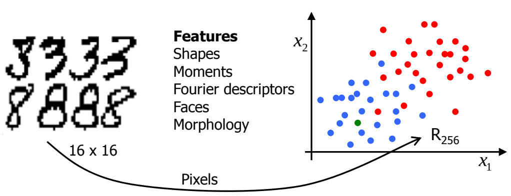
A

B

Nearest neighbours sorted:

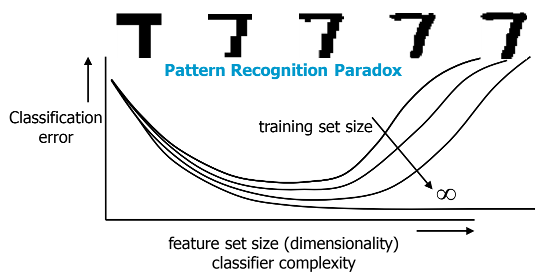
B A A A B A A B A B

Pixel Representation

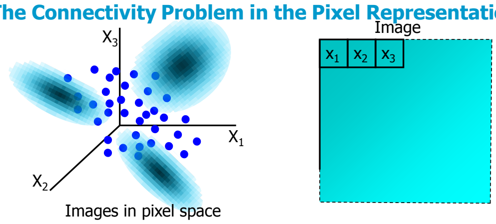


Pixels are more general, initially complete representation
Large datasets are available → good results for OCR

Peaking Phenomenon, Overtraining Curse of Dimensionality, Rao's Paradox

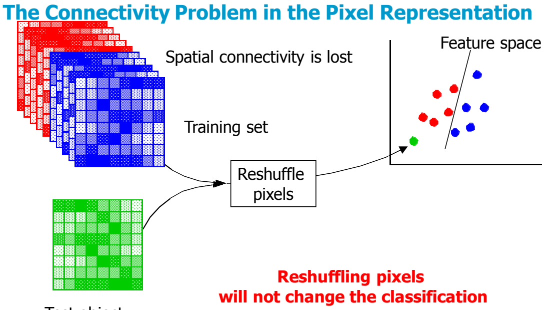


The Connectivity Problem in the Pixel Representation



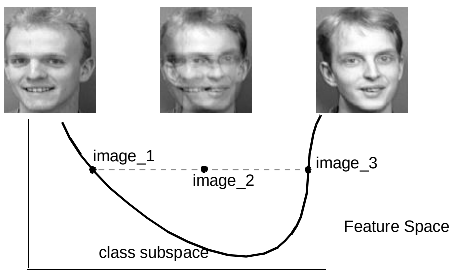
Dependent (connected) measurements are represented independently.
The dependency has to be refound from the data.

The Connectivity Problem in the Pixel Representation



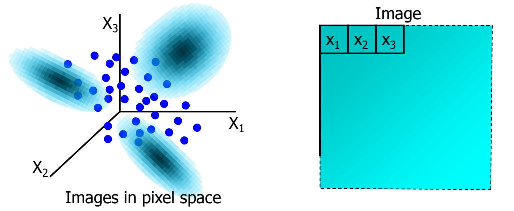
Reshuffling pixels
will not change the classification

The Connectivity Problem in the Pixel Representation



Interpolation does not yield valid objects

The Connectivity Problem in the Pixel Representation

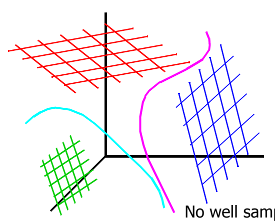


Dependent (connected) measurements are represented independently.
The dependency has to be refound from the data.

Reasons for selecting a pixel (sample) representation

- No good features can be found
- Sufficient training samples are available
- Direct, fast classification of the image
  (linear classifier == convolution)

Domains instead of Densities

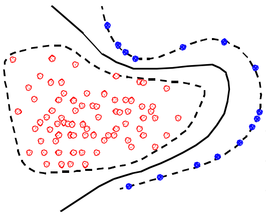


No well sampled training sets are needed.

Statistical classifiers have still to be developed.

Class structure ←→ Object invariants

Domain-based Classification

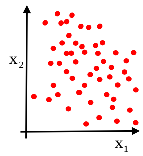


- Do not trust class densities.
- Estimate for each class a domain.
- Outlier sensitive.
- Distances instead of densities

How to construct domain based classifiers?

TU Delft

Wrong Intuition of High Dimensional Spaces



2D-intuition does not work for high dimensional spaces

All points are boundary points
1000 normally distr. points in R20: 95% on the convex hull.

Points tend to have equal distances
Squared Euclidean distances of points in R1024 are distributed as N(1024,32√2), so distances are all equal within 10%.

Class overlap is not visible
1000 points of two 5% overlapping classes in R50 can be linearly separable

Moreover:
do real world measurements with n > 100 really exist?
⇒ Subspace approaches

TU Delft

Reasons for selecting a pixel (sample) representation

- No good features can be found
- Sufficient training samples are available
- Direct, fast classification of the image
  (linear classifier == convolution)

TU Delft

Vector Representations

- Features: reduce → class overlap
- Pixels: dimensionality problems
- Dissimilarities?

TU Delft