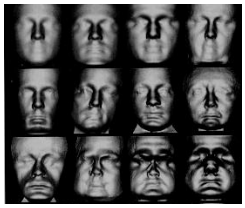
Unsupervised learning

Lecture 8

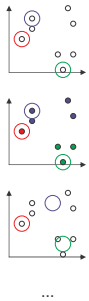


Outline

- Clustering
- Feature extraction
  - Review of algebra
  - Principal component analysis
    - Visualization of data
    - Creating new attributes
      - Example: Face recognition
  - Independent component analysis
    - Learning in the brain?

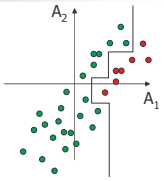
k-mean clustering

- Algorithm
  - Initially assign k-cluster centres to k randomly chosen instances
  - Repeat until converged
    - Assign each instance to the group that has the closest centre
    - Recalculate positions of the centres
- Implemented in Weka Explorer

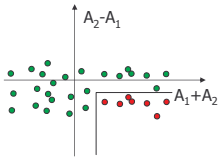


Need for feature extraction

- Learn training set using decision tree:



- The decision tree would be much simpler if the data were represented in new attributes: features



Review: Matrix multiplication

- Consider example:

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} 2 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1*2+0*1 & \\ & \end{bmatrix}$$

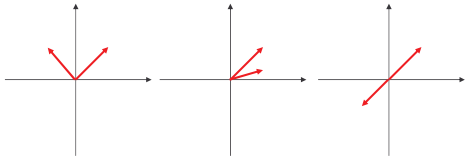
$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} 2 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1*2+0*0 \\ & \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} 2 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 1*2+2*1 & \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} 2 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 4 & 1*2+2*0 \end{bmatrix}$$

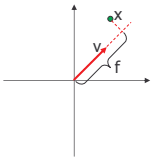
Review: bases

- A set of vector is a basis of a space
  - If any point in the space can be obtained by a linear superposition of basis vectors
- Examples



Review: bases…

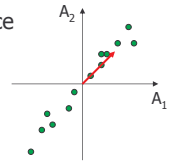
- Types of bases
  - Orthogonal – all basis vectors orthogonal
  - Orthonormal – all basis vectors orthogonal of length equal to 1
- Projection on a normal vector

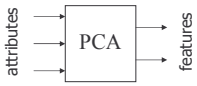


f=<x,v>

Principal component analysis

- Learning:
  - Finds directions in attribute space along which the data points have the largest variance
  - Vectors representing these directions are called Principal Components
- Feature extraction:
  - Project the data (vectors of attributes) on principal components





Finding principal components

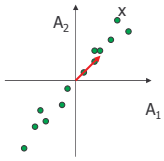
- Intuitively:
  - 1st PC is the direction in the space along which data is most spread
  - 2nd PC is direction in space orthogonal to the 1st PC along which data is most spread
  - Etc.
- Formally:
  - 1st PC is the eigenvector of covariance matrix corresponding to the highest eigenvalue
  - 2nd PC is the eigenvector of covariance matrix corresponding to the 2nd highest eigenvalue
  - Etc.
- Implemented in Weka Explorer

Projecting on principal components

- Consider data
- Principal components

$PC_1 = \left[1/\sqrt{2}, 1/\sqrt{2}\right]$

$PC_2 = \left[-1/\sqrt{2}, 1/\sqrt{2}\right]$



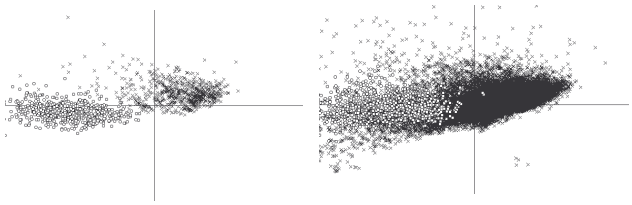
- Projection on principal components
  - E.g. $x = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$
  - Computing features: $PC \times x = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 0.7 \end{bmatrix}$
  - Representing data: $x = PC^T \begin{bmatrix} 3.5 \\ 0.7 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3.5 \\ 0.7 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

Properties of PCA

- Features extracted by PCA are uncorrelated
- First features carry most of information about data point
  - Note for those who know about eigenvalues:
    - The variance of data in the direction of a PC is equal to the eigenvalue corresponding to this PC
- Hence applications:
  - Visualizing data by plotting first features
  - Deriving new attributes:
    - Reducing number of attributes
    - Uncorrelated attributes -> Easier learning

PCA for data visualization

- Two first principal components from 40 dimensional data

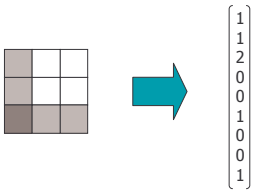


Healthy patients only

Both healthy and epileptic patients

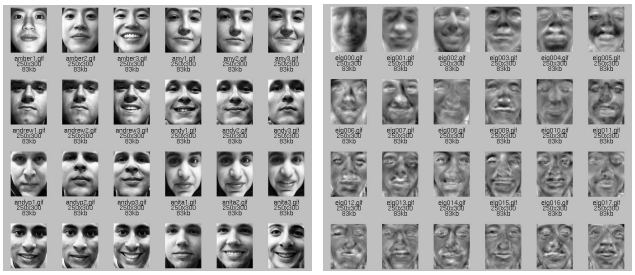
Representing images as vectors

- Vector created by concatenation of columns of pixels



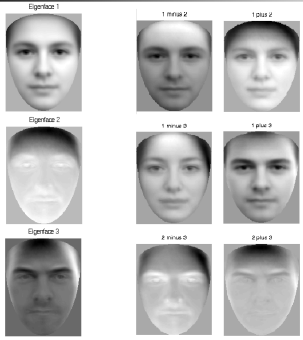
New attributes: eigenfaces

- Faces
- Principal components



http://www.owlnet.rice.edu/~elec301/Projects99/faces/images.html

Representing faces



- Eigenface 1
  - Main face feature
- Eigenface 2
  - Feminine features
- Eigenface 3
  - Masculine features

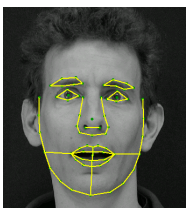
http://www.kent.ac.uk/physical-sciences/aog/facereco/

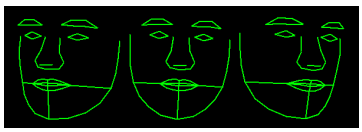
Importance of background knowledge

- Problem with eigenfaces:
  - If face is shifted by a few pixels: completely different features will be generated
    - This is easy to fix: e.g. centre the nose, scaled
  - If face is seen from a different angle: completely different features will be generated
    - This is very big problem for eigenface technique
- We can use background knowledge:
  - What really defines the face are not colours of pixels but distances between certain points on the face

Statistical shape model

- Positions of characteristic points can be extracted from the image
- PCA performed on co-ordinates of these points
- First feature describes rotation of the face

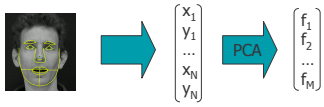




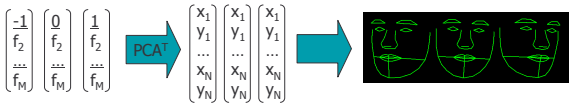
Tim Cootes
http://www.isbe.man.ac.uk/~bim/

Understanding features

- For each face:



- To see what does a feature represent: observe effect of varying the feature

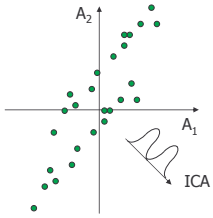


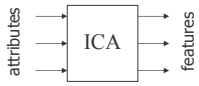
Example

- **M**odel-based **IN**terpretation of **OR**o-facial **I**mages, UCL
  - http://www.eastman.ucl.ac.uk/~dmi/MINORI/index.html
- The animation shows the second principal component varying between -3 and +3 standard deviations.

Independent component analysis

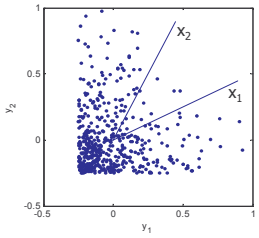
- Learning:
  - Finds independent features
  - Finds most interesting directions in attribute space
- Feature extraction:
  - Project the data (vectors of attributes) on independent components



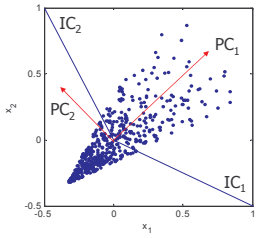


Example

- Independent underlying features



- Observable correlated data



Properties of PCA and ICA

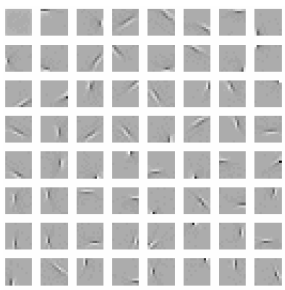
- When data Gaussian PCA & ICA give the same features
- PCA always generates orthogonal features while ICA does not need to
- Hence the hypothesis space of PCA is much smaller than that of ICA
  - Both are infinite, but
  - The number of free parameters estimated by PCA is much smaller than of ICA
  - Only use ICA if: number of training points >> number of attributes ^{2}

PCA vs. ICA

- PCA
  - Features uncorrelated
  - Information cumulated in first features
  - Fast learning always converging to unique solution
    - As searches smaller hypothesis space
- ICA
  - Features independent
  - Information more uniformly distributed among features
  - Slow iterative learning, that usually converges to different solutions each time is run
    - Local minima of independence

ICA - the way the brain learns?

- Features extracted from natural images using ICA remind "receptive fields" of neurons in the primary visual cortex



- But the brain probably uses nonlinear transformations

Summary

- Presenting data as vectors of meaningful features makes classification easier
- Techniques for feature extraction:

PCA	ICA
Uncorrelated features	Independent fetures
Cumulate information	Doesn't cumulate info.
Robust & fast	Local minima & slow

- Adding background knowledge to the feature extraction process produces really discriminative features