

Notes on Pattern Recognition

What is it?

image->feature extraction->feature vector->Classification->Class

What is feature extraction?

Convert a raw pattern into a feature vector.

Reduce redundancy in the pattern.

e.g. convert image to line drawing. Use techniques we have already seen e.g. edge detection, Fourier Features. Hough transform.

What is Classification

Assign the feature vector a class number

What is pattern space and feature space

Every possible image is a point in multidimensional space.

An image with 2 pixels is a point in 2D space.

An image with 1000*1000 pixels is a point in 1 million dimensional space.

The space for the raw image is called the pattern space

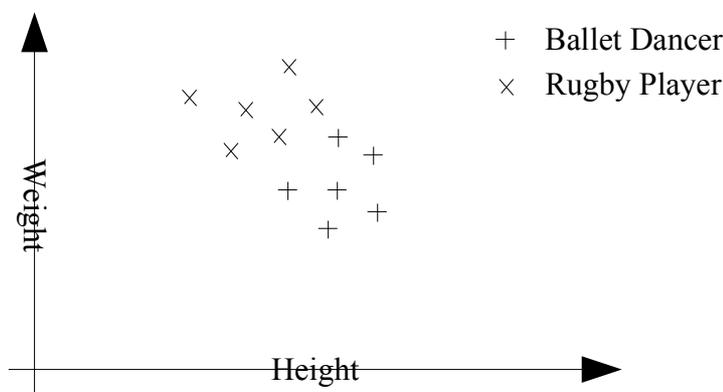
After feature extraction the pattern is in the feature space

Feature space will be smaller.

Example?

Ballet Dancers and Rugby Players

measure heights and weights of 12 people.



Pattern space (or feature space) must be partitioned through training.

Patterns from the same class must be close together (or the feature extraction is not good enough)

What is Training and Testing

The system is trained using a finite set of patterns- *the training set*.

If the correct classification for these patterns is known then this is **Supervised Learning**, otherwise it is **Unsupervised Learning**.

The system is evaluated using a different set of patterns - *the test set*.

How to describe a class

somehow we must describe the area in feature space occupied by a class, how:

parametric methods

use some parameters to describe the distribution e.g. assume gaussian and use mean, variance

e.g. Bayesian methods

e.g. average weight and height for a rugby player = (90,1.8) with variance=(5,.4)

discriminant functions

find a function which gives the decision boundary between classes

e.g, neural networks

e.g. 2 lines on diagram

structural methods

find rules for describing how patterns are created

e.g. syntactic methods

e.g. for a ballet dancer $h > 1.9$ and $w/h < 35$

What is Nearest Neighbour Classification?

create a discriminant function by remembering all training data

A pattern is classified as belonging to the class of the training pattern that is closest to it. To measure closeness use a **distance metric**.

For a feature vector $x = \{x_1, x_2, x_3, \dots, x_n\}$

and a training pattern $t = \{t_1, t_2, t_3, \dots, t_n\}$

Euclidean distance:

$$D^2 = \sum_i (x_i - t_i)^2$$

Dot Product Distance:

$$D = \sum_i (x_i * t_i)$$

Angle between vectors:

$$D = \sum_i (x_i * t_i) / (|x_i| * |t_i|)$$

Efficiency

Training is trivial, just store all patterns, this may require a lot of storage.

Classification may be time consuming since all stored patterns must be compared.

Optimality

Nearest Neighbour is not optimal. Making simple assumptions it can be shown that the error probability is bounded by twice the optimal error.

Nearest neighbour classification is prone to errors due to rogue patterns. A rogue pattern is a mislabelled pattern and causes classification errors.

What is K Nearest Neighbour?

To eliminate the problem caused by rogue patterns use not just the nearest neighbour but a group of them.

Using K neighbours, take a majority vote of their classes to give a classification.

Optimality

As K gets larger this method approaches the optimal decision.

Speeding up nearest neighbour methods.

The biggest problem with this method is the time it takes to calculate the distances to the training examples.

Possible solutions are:

- Only store examples near the decision boundary.

- Use a pre computed search tree and branch and bound to search for the nearest neighbour.

What is Committee method

combine lots of classifiers to get better result.

How does a real committee work

makes good decision if

- all members are independent (not always true) and all contribute equally (rarely true)

independent?

- use different data (not read same books)

- or have same data but use different methods (e.g. statistician, marxist, postmodernist, on committee will make interesting decisions)

So how does CM work?

- use different training data for each classifier

- or use very different methods on same data

- just add up all the classifications, most votes wins

- also gives a confidence (if winner is close to second then not very confident)

What is Unsupervised Learning?

Sometimes we don't have correct classifications for training patterns.

What is clustering

use the fact that similar patterns group together (in clusters) to assign classes to them.

Clustering algorithms perform dimensionality reduction. They do this because they take a high dimensional pattern space and produce a lower dimensional space.

Hopefully the lower dimensional space still contains most of the information from the original space.

When it is used

- We have no labels for classes or we want to find inherent properties of data.

- We know the number of classes.

- We know very little about the process which has created the patterns.

What is the K means algorithm:

Assume that there are K clusters in the data, i.e. K classes.

- choose small random values for each cluster mean

- $m_1(0), m_2(0), m_3(0), \dots, m_K(0)$ where $m_i(0) = (x_1, x_2, \dots, x_n)$

- Use the training set to assign a cluster to each training pattern.

Given a training pattern y , y is in cluster i at time t - $C_i(t)$ if
Forall j , $d(y, m_i(t)) = \min(d(y, m_j(t)))$
where d is some distance metric.
i.e. y is in cluster i , if i has the closest mean.
Calculate new values for $m_1(t) \dots m_K(t)$
 $m_i(t+1)$ minimises $\text{Sum}(d(y, m_i(t)))$ where y is in $C_i(t)$
For the Euclidean distance this is just the mean of all the patterns assigned to class i .
Repeat from 2 until

Forall i , $m_i(t+1) - m_i(t) < \epsilon$
i.e. the clusters are not moving much.

What is ISODATA

An algorithm based on K means that does not need to know the number of clusters.
It starts with one cluster and splits/merges clusters according to certain parameters.

What parameters does it need?

Minimum distance between two cluster centres below which they are merged.
Maximum standard deviation for a cluster above which it is split into two clusters.
Minimum proportion of training patterns in a cluster.
Amongst many others.

How does it work?

Modify the K means algorithm we need to add an extra step 4 which can split or merge clusters.