Introduction of Machine Learning (part II: examples of classification)

Ekberg and Maki

September, 2013

DD2431, CSC/KTH

In this part we will visit:

- Some more examples
- Concept of classification
  - Hand-written digit recognition
- Simple approaches for classification
  - Nearest Neigbour method

Where is machine learning useful?

- A pattern exists.
- Data available for training.
- Hard/impossible to define rules mathematically.

Related terms on data analysis

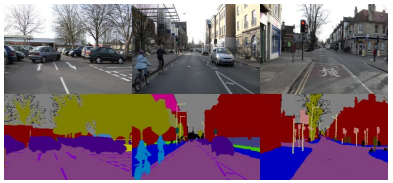
- Pattern Recognition
- Data Mining
- Statistics

Examples of applications

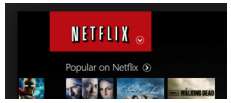
- Optic character recognition (OCR)
- Medicine
- DNA analysis
- Remote sensing
- Speech Technology

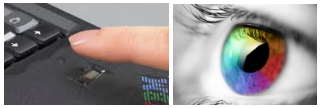
- Computer Vision
- Robotics
- Finance
- Recommender systems: books, movies
- Biometrics: fingerprint, iris, face

…



(The Cambridge-driving Labeled Video Database)





Classification

- We would like to enable a computer to learn from data to answer a question - "What is it?"

  You're given sample data (for finding patterns).

The framework of classification

- Training phase: to give the concept of classes to a machine using labeled data
- Testing phase: to determine the class of new unseen (unlabeled) data

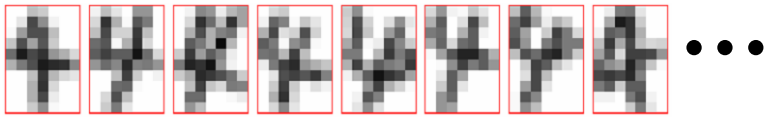
Example: Hand-written digits

One of the first commercial system with ML, used for zip codes

Training samples:

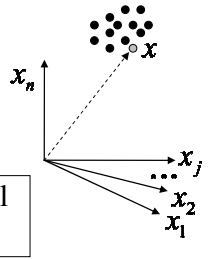
Feature extraction

Pattern vectors: normalized & blurred patterns



Feature extraction for digit recognition

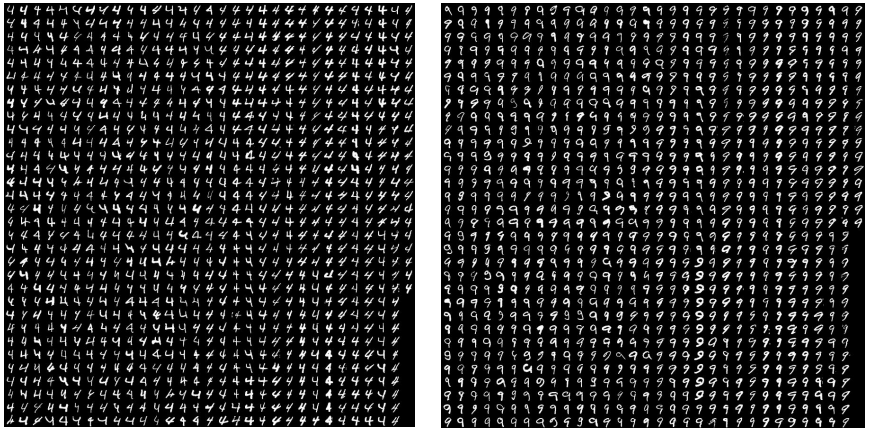
- Represent an image by a feature vector
  - Sampling: n = 7 x 10 pixels



  - A set of n gray values: $x = (x_1, \ldots, x_n)$

    i.e. corresponding to a point in feature space

More training samples of “4” and others



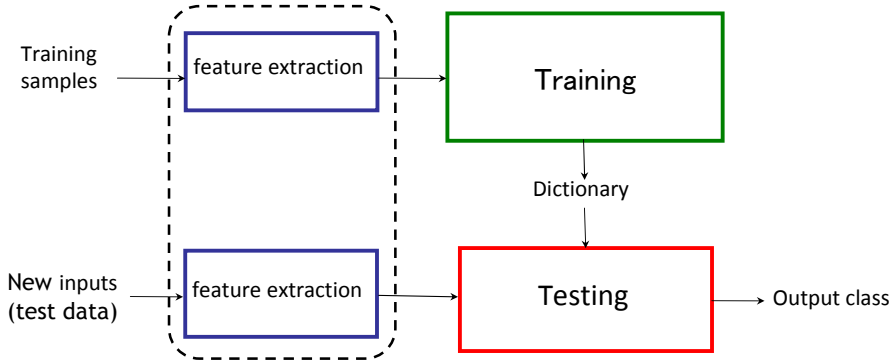
http://yann.lecun.com/exdb/mnist/

Example: Face images

Training samples of frontal faces

(Face image database, CMU)

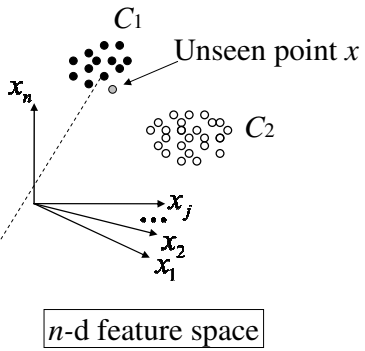
Schematic of classification



Nearest Neighbour methods

- Binary classification
  - N_1 samples of class C_1
  - N_2 samples of class C_2
  - Unseen data x

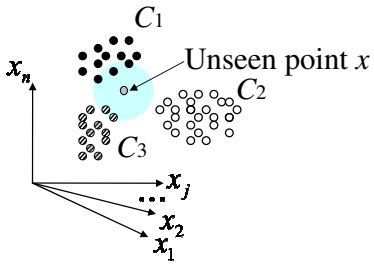
  → Compute distances to $N_1 + N_2$ samples



- Find the nearest neighbour

→ classify x to the same class

- k-nearest neighbour rule
  - Compute the distances to all the samples from new data x
  - Pick k neighbours that are nearest to x

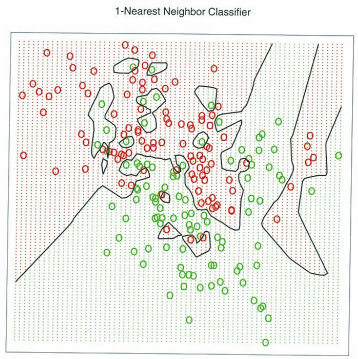


→ Majority vote to classify point x (Nearest Neighbour is 1-NN)

- How does k-NN compare to 1-NN ?

What is the influence of k ?

$k = 1$ $\qquad$ $k = ?$

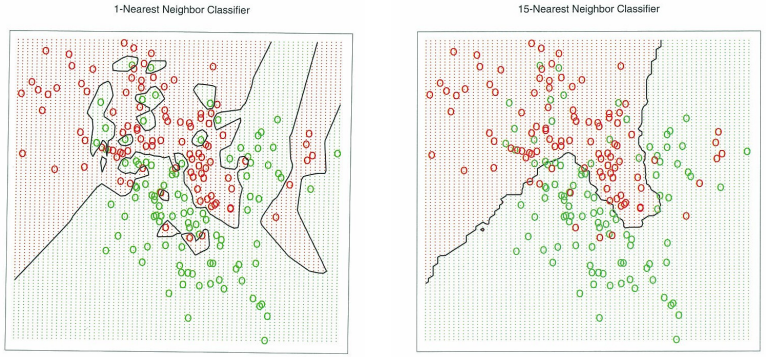


?

Decision boundaries with different k

$k = 1$ $\qquad$ $k = 15$

No misclassifications on training data $\qquad$ More generalized



(T. Hastie et al, The Elements of Statistical Learning)

Pros and cons of k-NN

- k-NN / 1-NN comparison summary
  - the boundary becomes smoother as k increases
  - lower computational cost for lower k
  - k-NN better generalizes given many samples
- Pros:
  - simple; only with a single parameter k
  - applicable to multi-class problems
  - good performance, effective in low dimension data
- Cons:
  - costly to compute distances to search for the nearest
  - memory requirement: must store all the training set

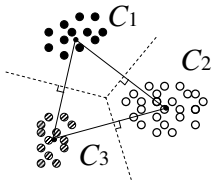
Notes on "generalization"

- Our goal is to determine the class of unseen data.
- Strategies/parameters that achieve minimum loss on training samples is not necessarily best for test data.
- We want the machine to learn the true pattern (and not noise) that resides in the sample data for generalization.

Discriminant function

- Remember all the samples?
  - Simply used all the training data in k-NN …
  - Still cover only a small portion of possible patterns
- Define a class by a few representative patterns
  - e.g. the centroid of class distribution

Extreme case: one vector per class



Formulation: one prototype per class

- K classes: $C^{(1)},\cdots,C^{(K)}$
- K prototypes: $a^{(1)},\cdots,a^{(K)}$

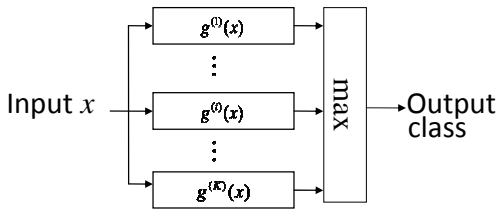
Consider Euclidean distances between the new input x and the prototypes:

$$\| x-a^{(i)} \|^2 = \| x \|^2 - 2a^{(i)T}x + \| a^{(i)} \|^2$$

→ Choose the class that minimises the distance.

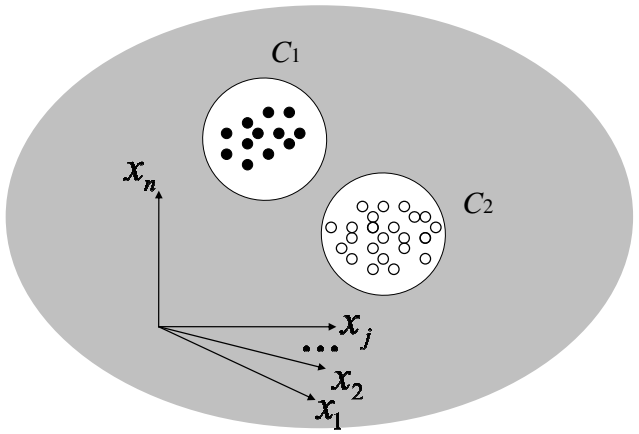
Discriminant function

$$g^{(i)}(x) \equiv a^{(i)T}x - \frac{1}{2}\| a^{(i)} \|^2$$



Setting the "don't know" category

- Reject if the distance is above the threshold

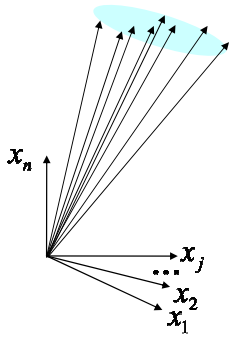


Subspace Method

- Exploit localization of pattern distributions

Samples in the same class such as a digit (or face images of a person) are similar to each other.
They are localized in a *subspace* spanned by a set of basis u_i .

u_i : reference vectors (orthogonal basis)



OCR system (a historical example)

Recognition of letters; 2000 alphanumeric chars/sec., 200 sheets/min. Analog circuit for similarity calculation.

Face Recognition (a biometric example)

Security system

FacePass(R)

Recognition while walking

SmartConcierge(R), 2007

Keywords to remember

- Classification
  - Feature extraction
  - Training/Testing
  - Generalization
- Classification methods
  - Nearest Neighbour rule
- Decision trees (to come)