Lecture 1: Introduction
DD2421 Machine Learning

Part II: Classification problems and Nearest Neighbor rule

Atsuto Maki

Autumn, 2020

In this part we will visit:

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

Where is machine learning useful?

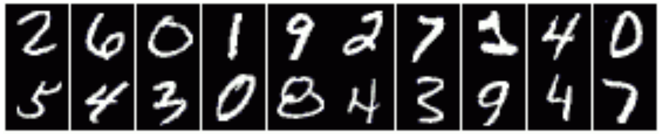
- A pattern exists.
- Data available for training.
- Hard/impossible to define rules mathematically.

Related terms on data analysis

- Pattern Recognition
- Data Mining
- Statistics

ML applications are pervasive

- Optic character recognition (OCR)



- Medical diagnosis
- DNA analysis

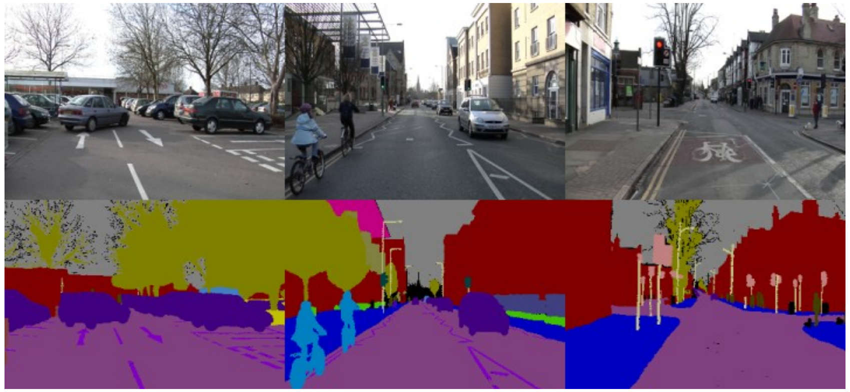
(by bejarprints)

- Remote sensing
- Speech technology

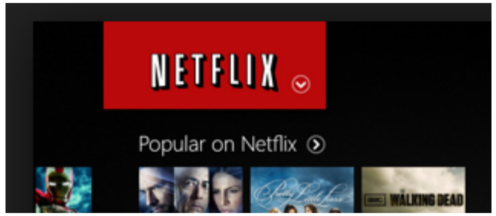
- NLP / automatic translation

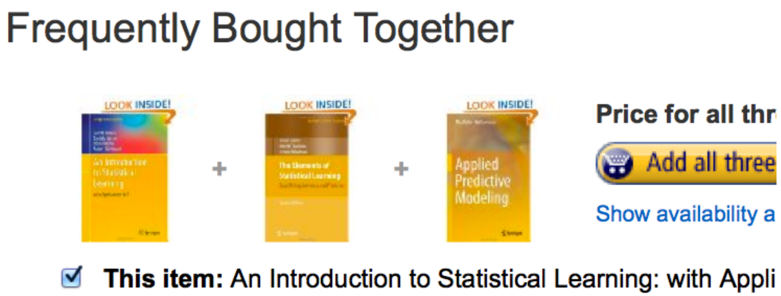
- Computer Vision
- Robotics
- Finance
- Recommender systems: books, movies
- Biometrics: fingerprint, iris, face

…



(The Cambridge-driving Labeled Video Database)





Concept of 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).

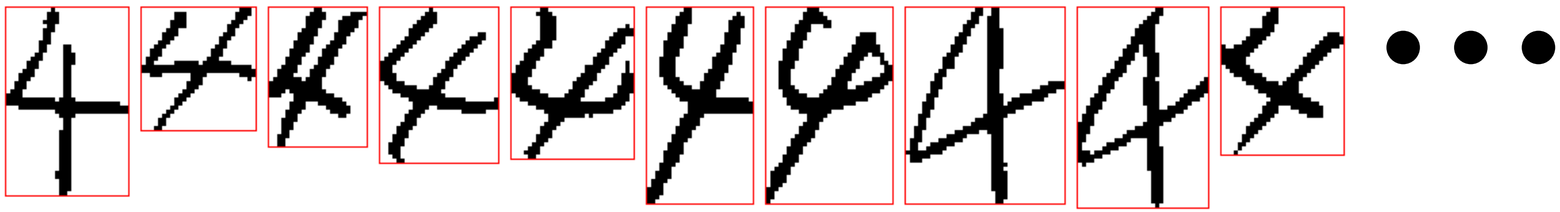
The framework of classification

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

Example: Hand-written digits

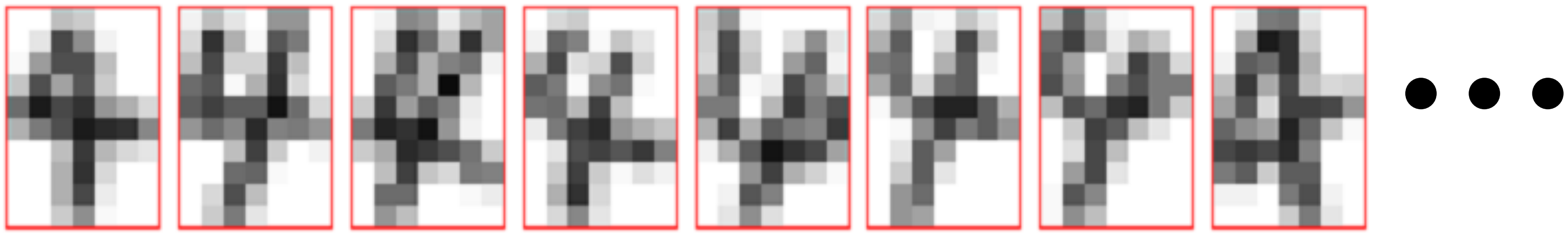
One of the first commercial system with ML, used for zip codes

Training samples:



Feature extraction

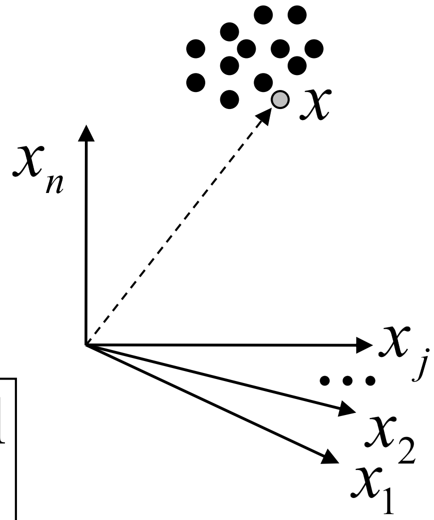
Pattern vectors: normalized & blurred patterns



Feature extraction for digit recognition

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

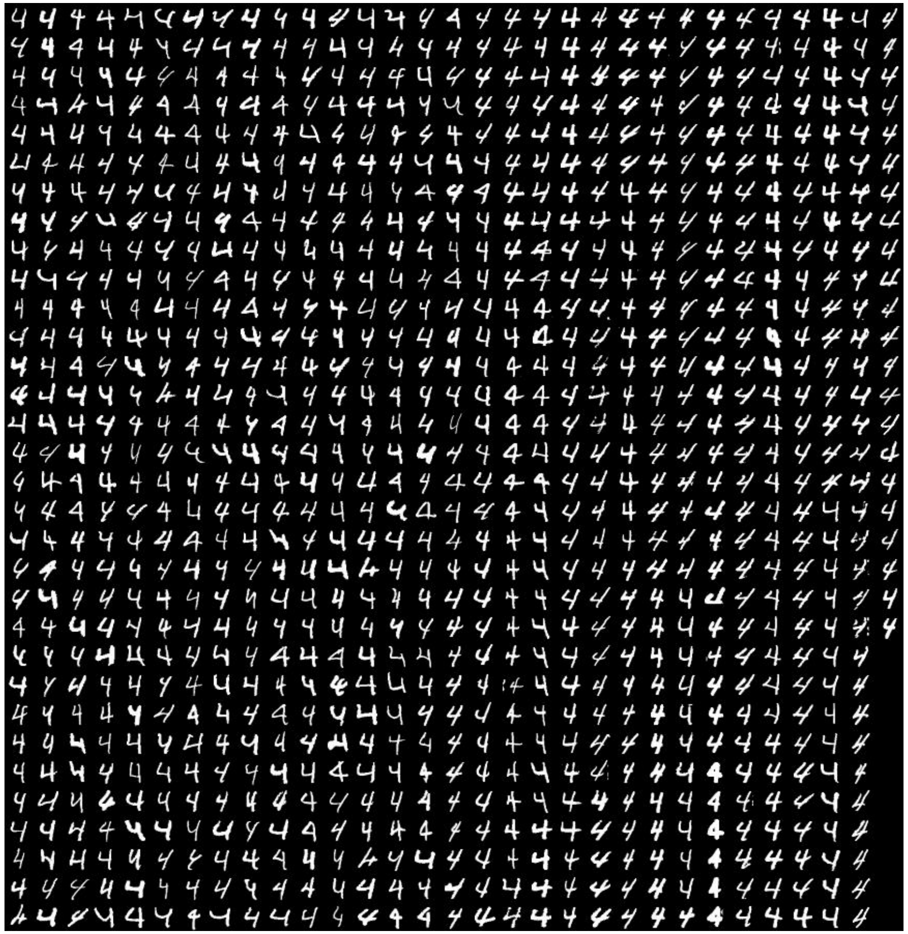
n-dimensional feature space

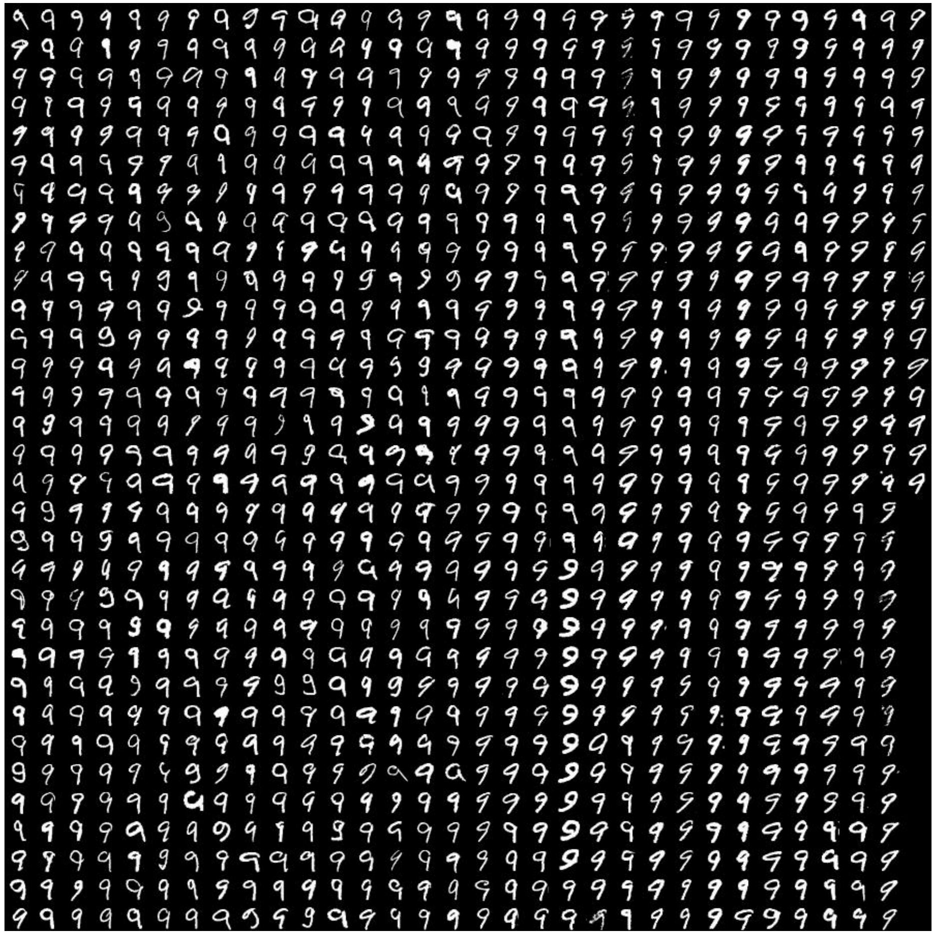


  - A set of n gray values: $\mathrm{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)$

  i.e. corresponding to a point in feature space

More training samples of “4” and others





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

OCR system (a historical example)

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

http://museum.ipsj.or.jp/en/heritage/ASPET/71.html

Example: Face images

Training samples of frontal faces

(Face image database, CMU)

Face Recognition (a biometric example)

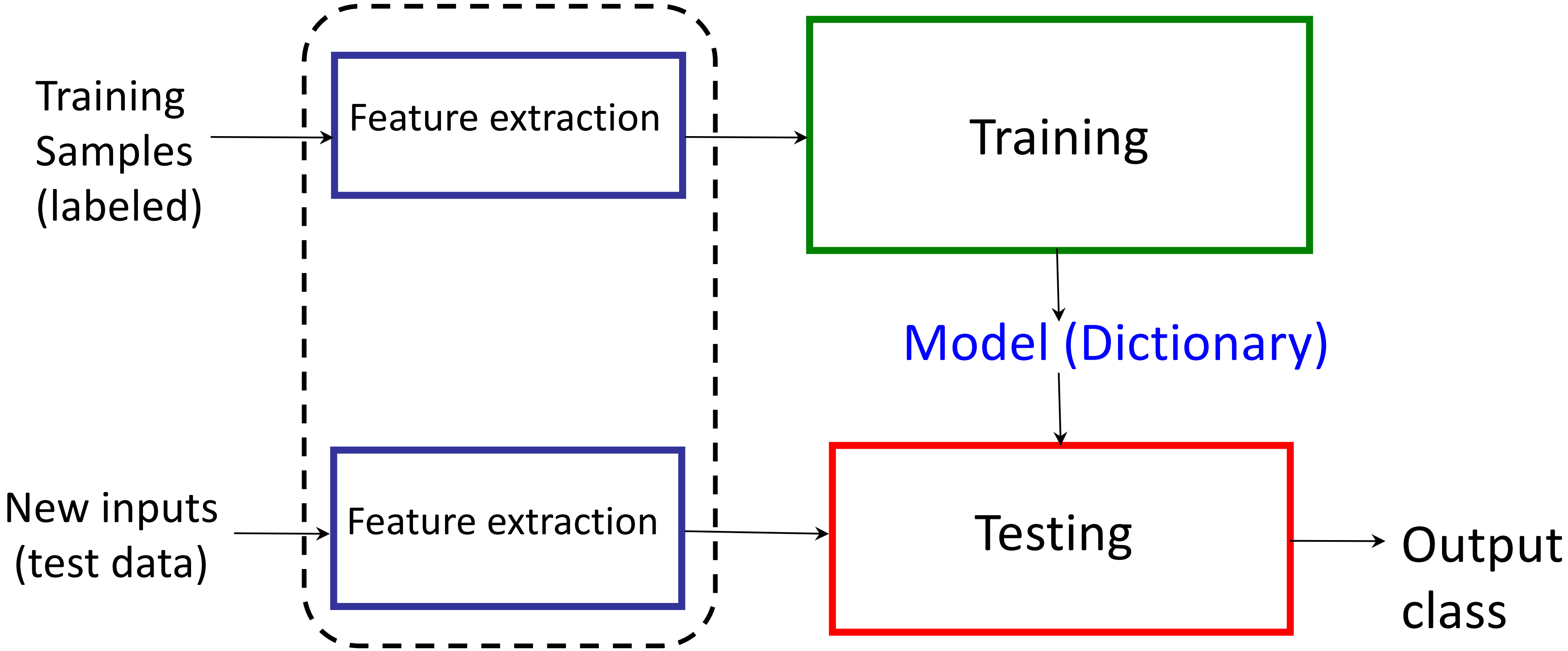
Security system

FacePass(R)

Recognition while walking

SmartConcierge(R), 2007

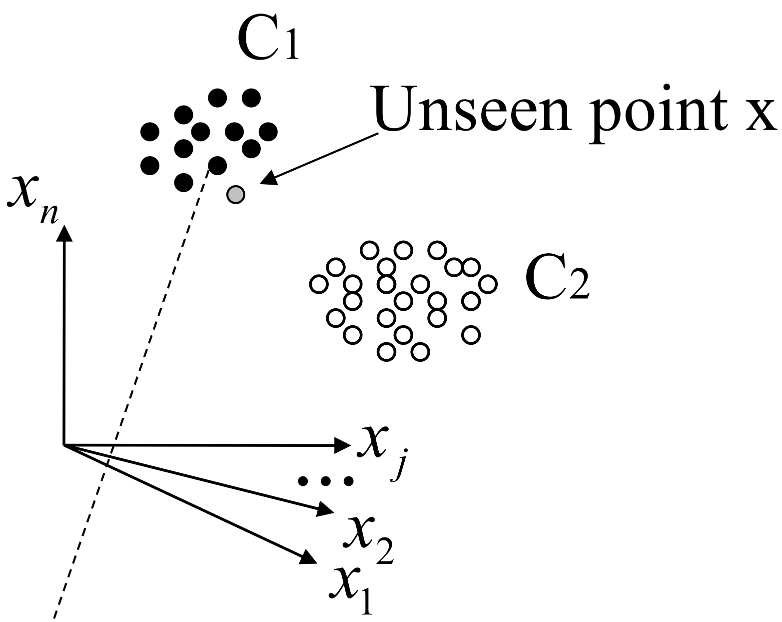
Schematic of classification



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 all the $N_1 + N_2$ samples



n-d feature space

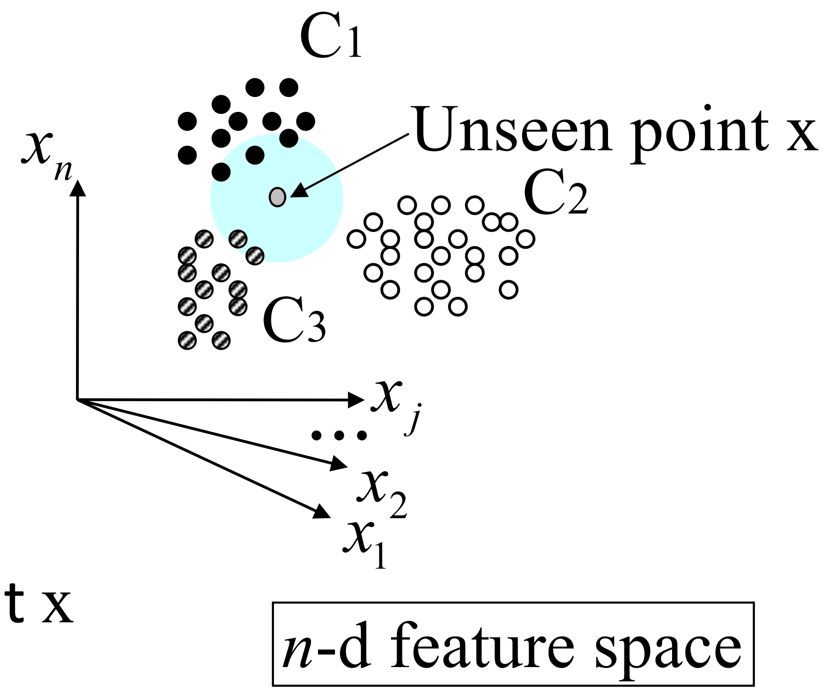
- Find the nearest neighbour

→ 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)

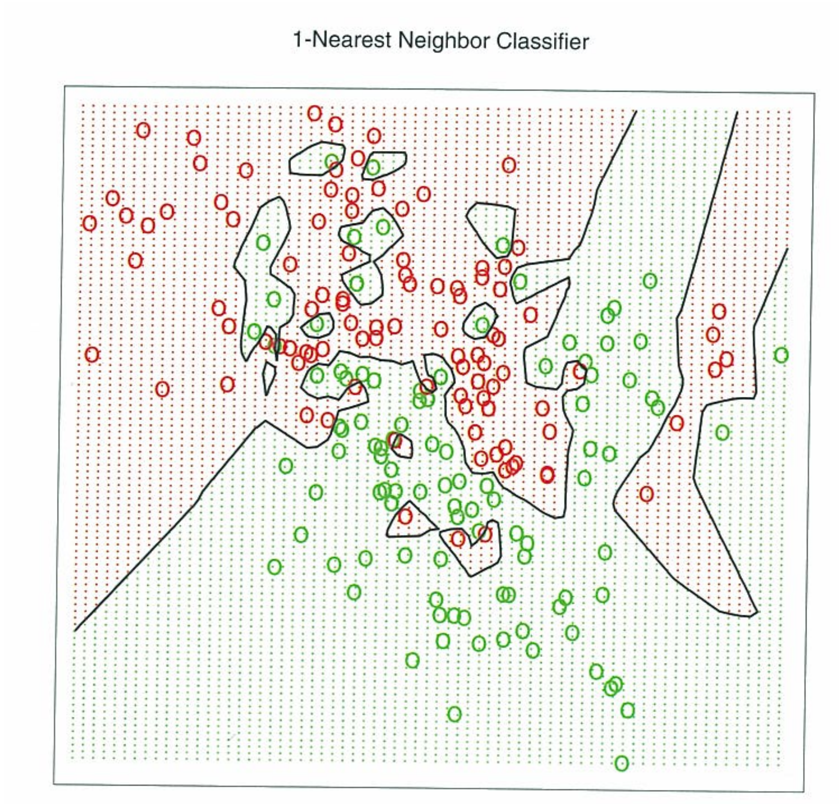


- How does k-NN compare to 1-NN ?

What is the influence of k ?

$k = 1$

$k = ?$

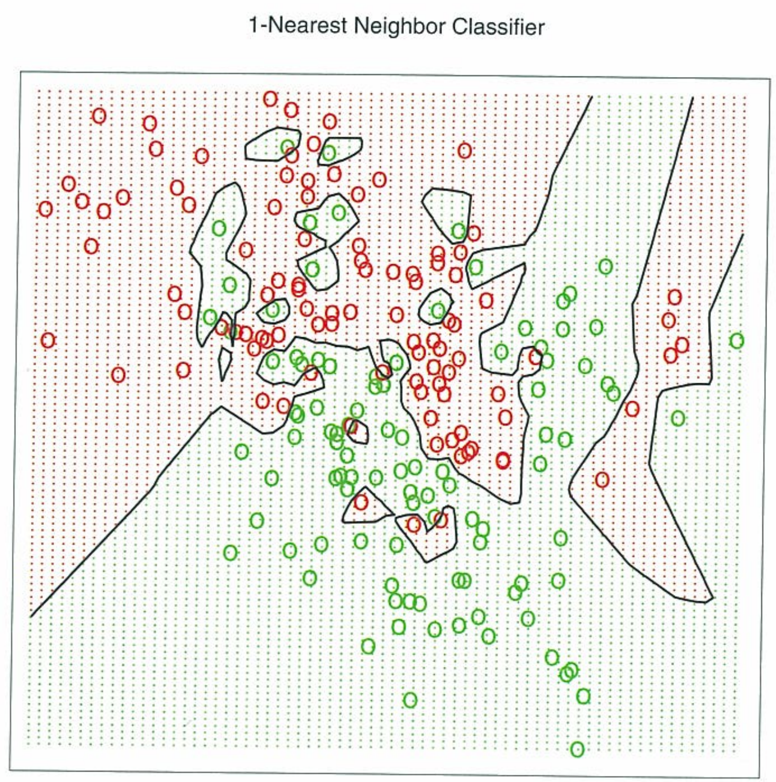


?

Decision boundaries with different k

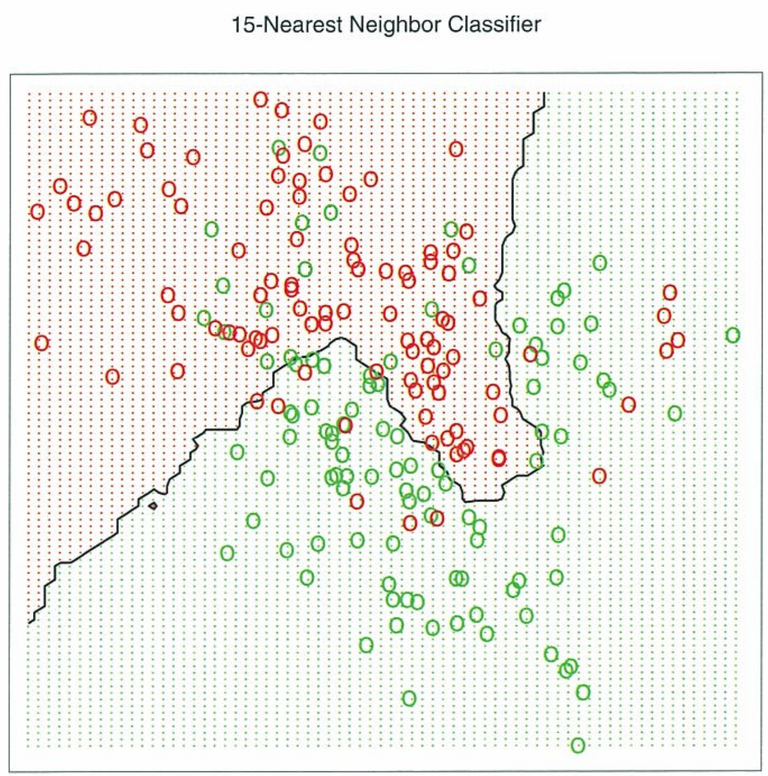
$k = 1$

No misclassifications on training data



$k = 15$

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.

Keywords to remember

- Classification
  - Feature extraction
  - Training/Testing
  - Generalization

- Classification methods
  - Nearest Neighbour rule

- Decision trees (to come)