Data Mining

Semi-supervised classification

Nearest Neighbors

24/11/2021

NOVA-IMS

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

AGENDA

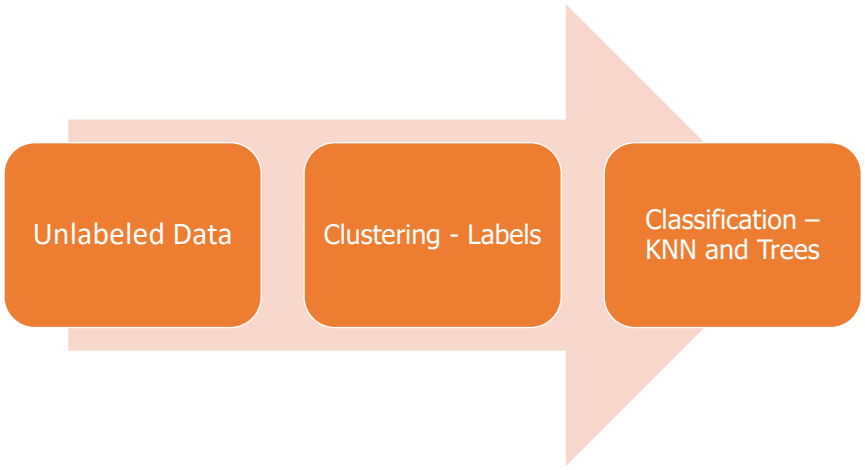
- Cluster analysis
  - Semi-supervised classification
    - Nearest Neighbors

2

Semi-supervised classification

3

Going from clustering into classification



4

Going from clustering into classification

5

k-nearest neighbors (k-NN)

6

k-nearest neighbors

- Instance based classification:
  - Simplest form of learning;
  - Training instances are searched for instances that most closely resembles new instance;
  - The instances themselves represent the knowledge;
  - Also called instance-based learning
  - Similarity function defines what's "learned"

k-nearest neighbors

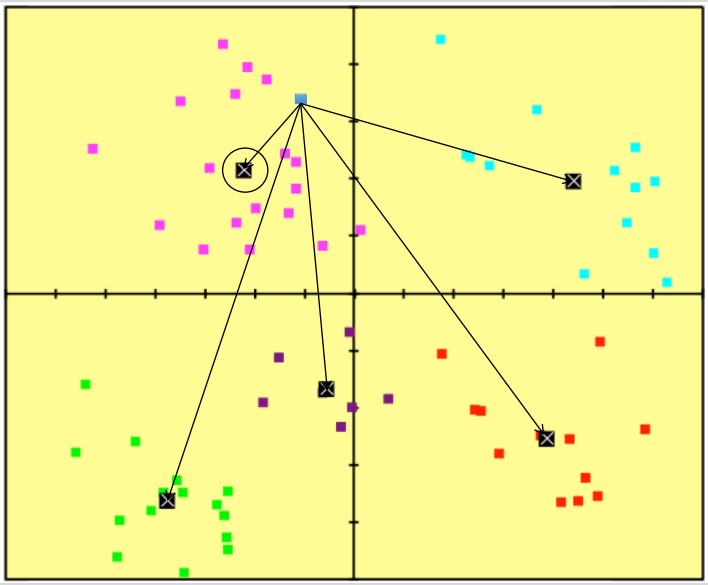
- Requires three things:
  1. The set of stored records (with labels)
  2. A distance metric to compute distance between records (can use Euclidean distance)
  3. The value of k, the number of nearest neighbors to retrieve

k-nearest neighbors

- To classify an unknown record:
  1. Compute distance to other training records
  2. Identify k nearest neighbors
  3. Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

9

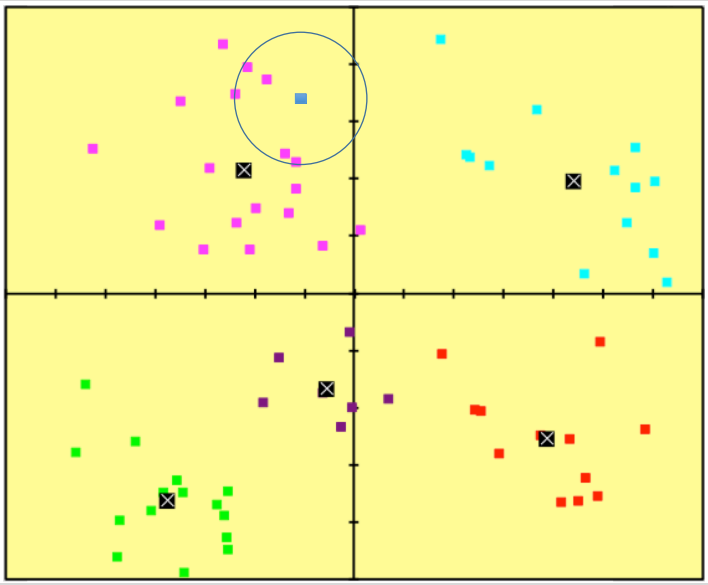
Going from clustering into classification



10

NOVA IMS Information Management School

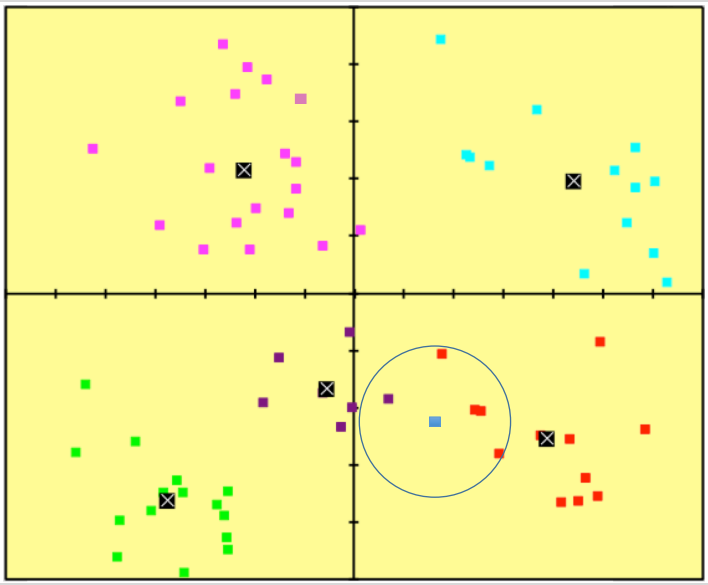
Going from clustering into classification



11

NOVA IMS Information Management School

Going from clustering into classification



12

k-nearest neighbors

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

  - Determine the class from nearest neighbor list
    - Take the majority vote of class labels among the k-nearest neighbors
    - Weigh the vote acording to distance
      - weight factor, w = 1/d2

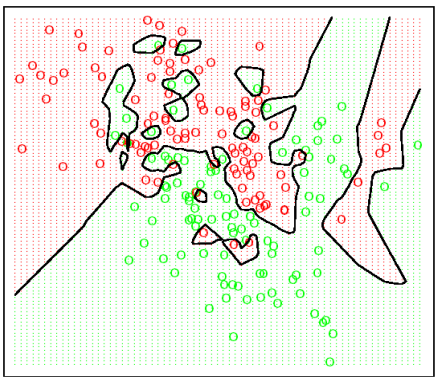
k-nearest neighbors

- k-nn frontiers (and the number *k*):
  - Large k
    - Smooth frontiers
    - Unable to detect small variations
  - Small k
    - Very sensitive to outliers
    - Crisp frontiers

NOVA IMS Information Management School

k-nearest neighbors

1-nn



15-nn