K-Nearest-Neighbours

Instance-based representation and learning

Lecture 5 – Part 2

Marina Santini

Acknowledgements
Slides borrowed and adapted from:
Data Mining by I. H. Witten, E. Frank and M. A. Hall

Outline

- Instance-based representation
- Instance-based learning
- Discussion

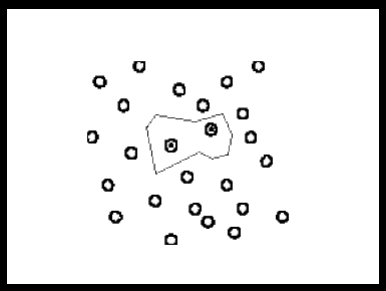
Instance-based representation

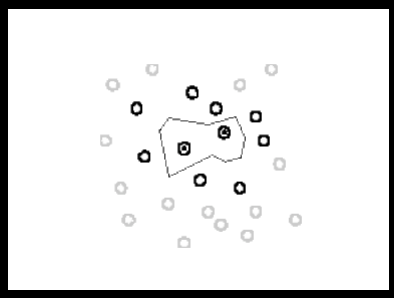
- Simplest form of learning: *instance-based* learning
  - Training instances are searched for instance that most closely resembles new instance
  - The instances themselves represent the knowledge
- Similarity function defines what's "learned"
- Instance-based learning is *lazy* learning
- Methods: *nearest-neighbor, k-nearest-neighbor, …*

The distance function

- Simplest case: one numeric attribute
  - Distance is the difference between the two attribute values involved (or a function thereof)
- Several numeric attributes: normally, Euclidean distance is used and attributes are normalized
- Nominal attributes: distance is set to 1 if values are different, 0 if they are equal
- Are all attributes equally important?
  - Weighting the attributes might be necessary

Learning prototypes





- Only those instances involved in a decision need to be stored
- Noisy instances should be filtered out
- Idea: only use *prototypical* examples

Instance-based learning

- Distance function defines what's learned
- Most instance-based schemes use *Euclidean distance*:

$$\sqrt{(a_1^{(1)}-a_1^{(2)})^2+(a_2^{(1)}-a_2^{(2)})^2+\ldots(a_k^{(1)}-a_k^{(2)})^2}$$

- $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$: two instances with k attributes
- Taking the square root is not required when comparing distances
- Other popular metric: *city-block metric*
  - Adds differences without squaring them

Normalization and other issues

- Different attributes are measured on different scales ⇒ need to be *normalized*:

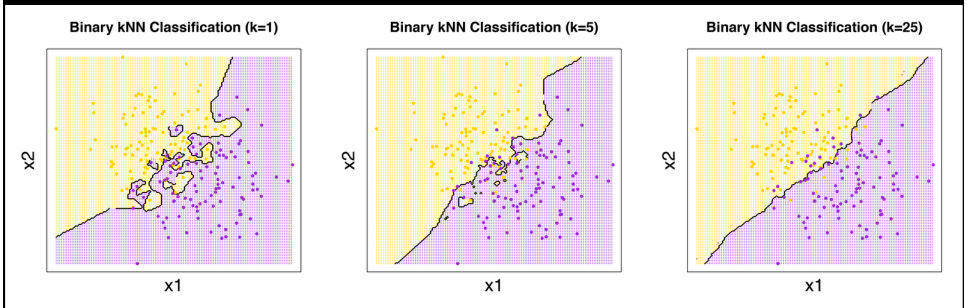
$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

v_i : the actual value of attribute i

- Nominal attributes: distance either 0 or 1
- Common policy for missing values: assumed to be maximally distant (given normalized attributes)

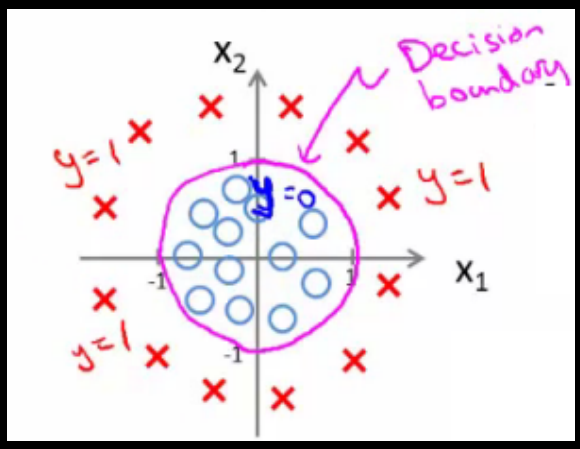
How many neighbours?

k-neighbours



Decision boundary

The boundary



Discussion of nearest-neighbor learning

- Often very accurate
- Assumes all attributes are equally important
  - Remedy: attribute selection or weights
- Possible remedies against noisy instances:
  - Take a majority vote over the *k* nearest neighbors
  - Removing noisy instances from dataset (difficult!)
- Statisticians have used *k*-NN since early 1950s
  - If $n \rightarrow \infty$ and $k/n \rightarrow 0$, error approaches minimum

More discussion

- Instead of storing all training instances, compress them into regions
- Another simple technique (Voting Feature Intervals):
  - Construct intervals for each attribute
    - Discretize numeric attributes
    - Treat each value of a nominal attribute as an “interval”
  - Count number of times class occurs in interval
  - Prediction is generated by letting intervals vote (those that contain the test instance)

The end