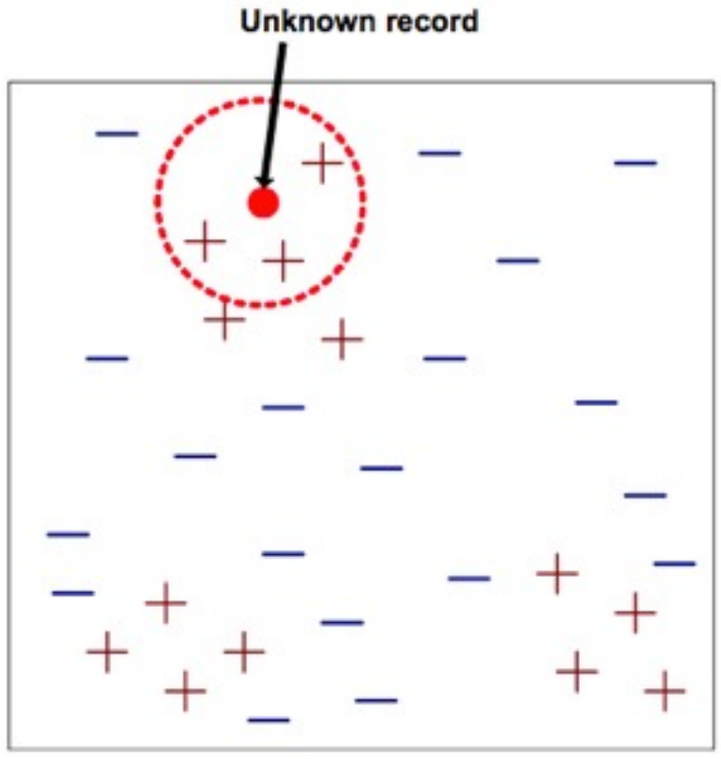
Data Mining
Classification: Alternative Techniques

Mauro Sozio

slides adapted from “Introduction to Data Mining” by Tan, Steinbach, Kumar.

K-Nearest Neighbor Classifier

- Training:
  - turn dataset into vectors (e.g. points euclidean space)
  - load them into main memory
- Prediction
  - find the k nearest points
  - output majority class in those points
- Requires
  - distance function
  - a value of k

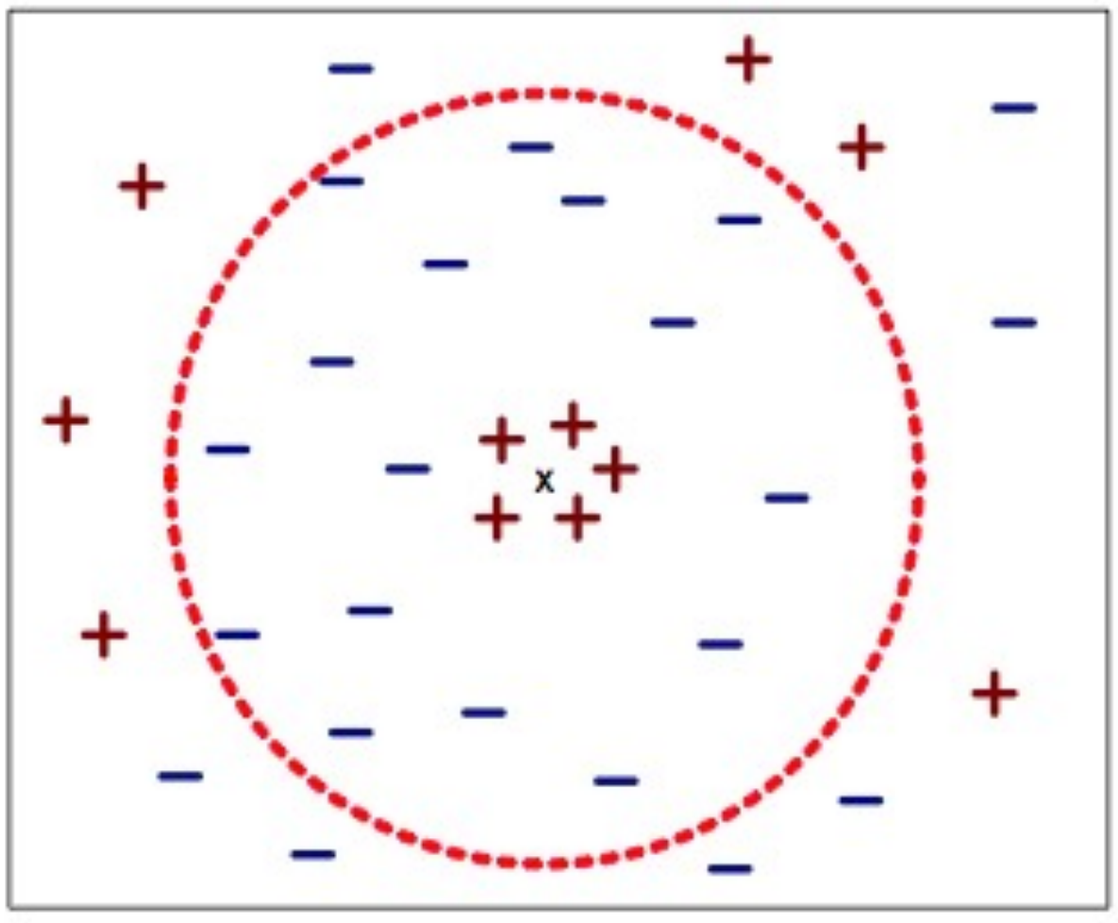


K-Nearest Neighbor Classifier

- Distance functions: euclidean distance but also cosine similarity
- Prediction: other strategies are possible (e.g. weighted vote according to the distances).

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



Nearest Neighbor Classification…

- Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

- It suffers from the curse of dimensionality (scalability issues, data becomes too sparse)

Nearest neighbor Classification…

- k-NN classifiers are lazy learners
  - do not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive