Data Mining
Classification: Alternative Techniques

Lecture Notes for Chapter 4

Instance-Based Learning

Introduction to Data Mining , 2nd Edition

by

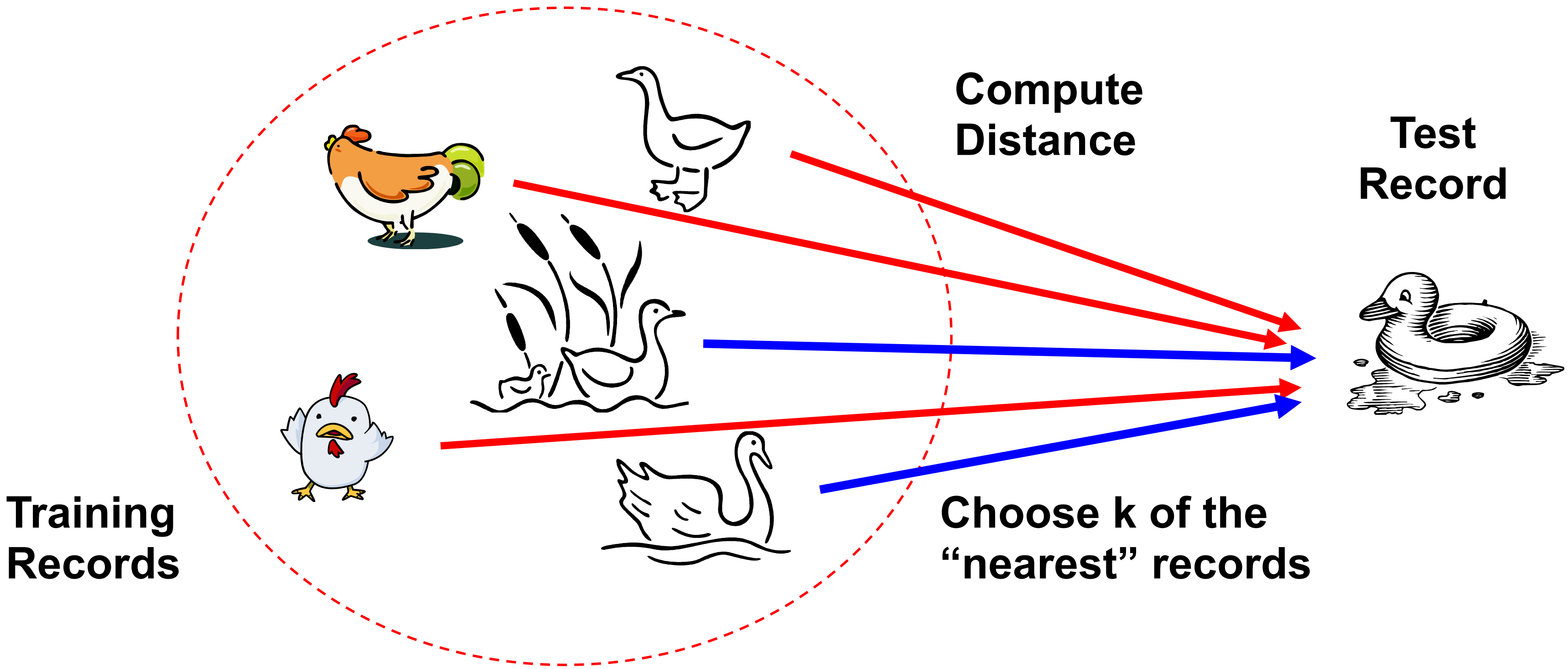
Tan, Steinbach, Karpatne, Kumar

Instance Based Classifiers

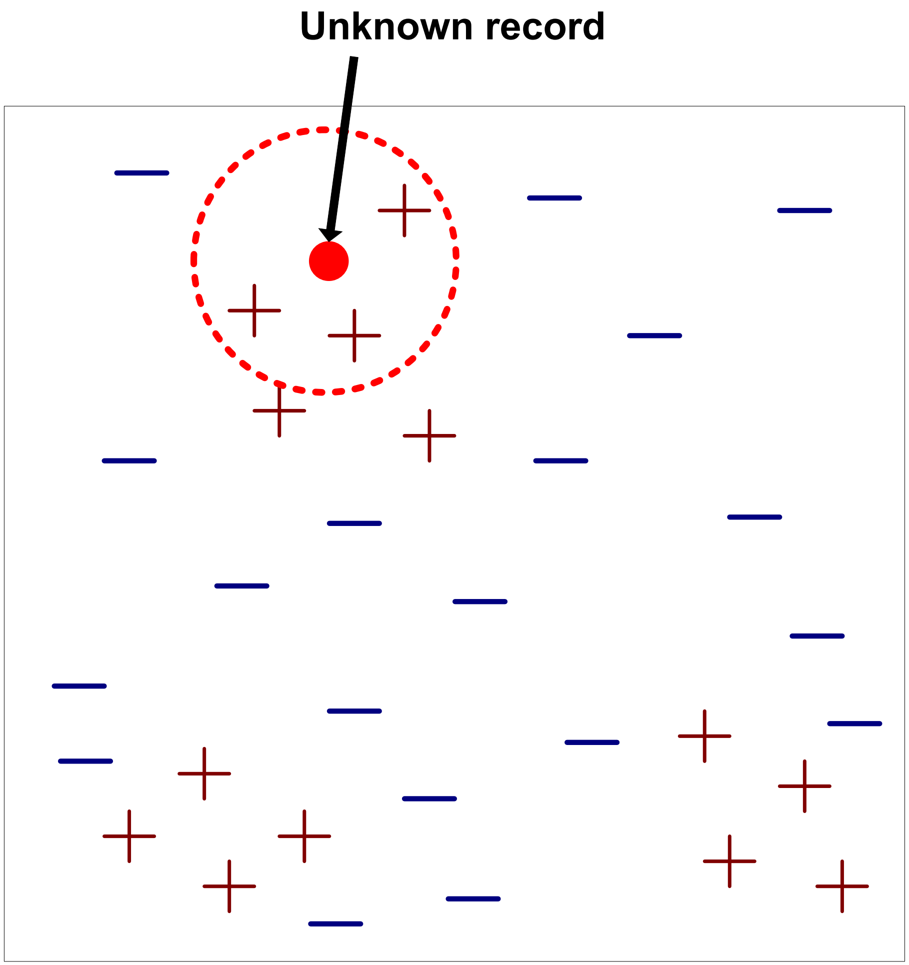
- Examples:
  - Rote-learner
    - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
  - Nearest neighbor
    - Uses k "closest" points (nearest neighbors) for performing classification

Nearest Neighbor Classifiers

- Basic idea:
  – If it walks like a duck, quacks like a duck, then it's probably a duck



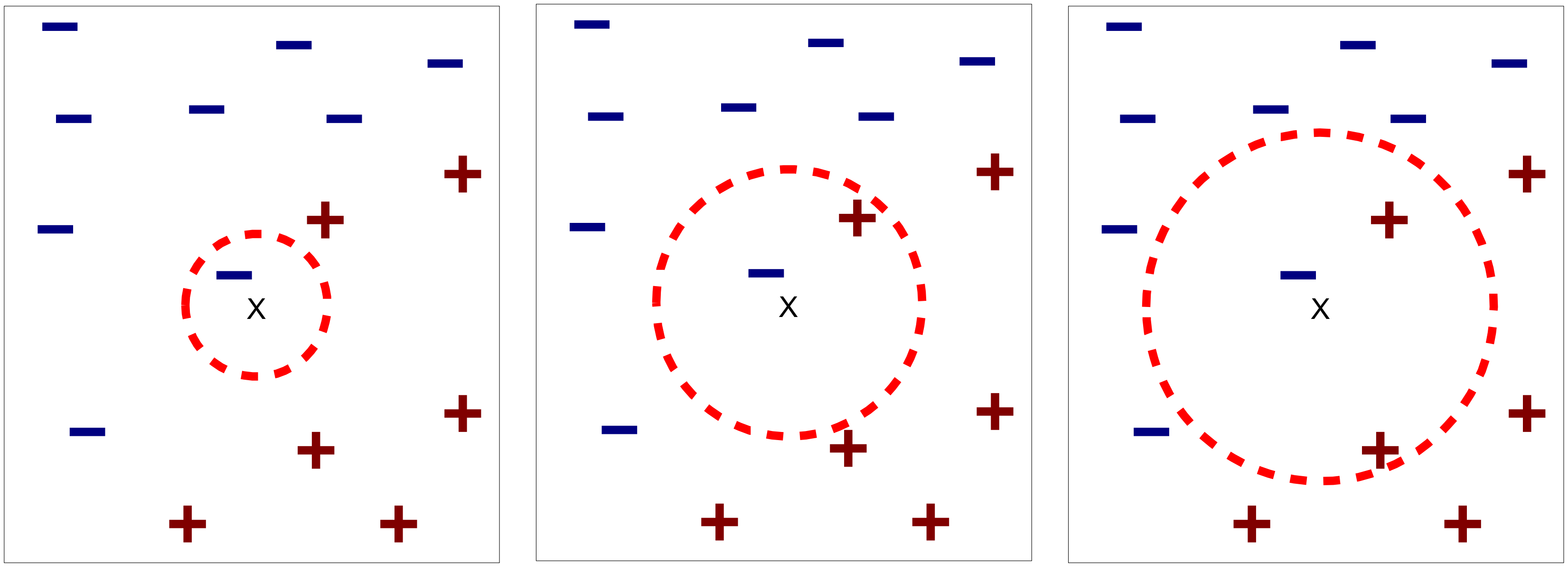
Nearest-Neighbor Classifiers



- Requires three things
  - The set of labeled records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

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

Definition of Nearest Neighbor

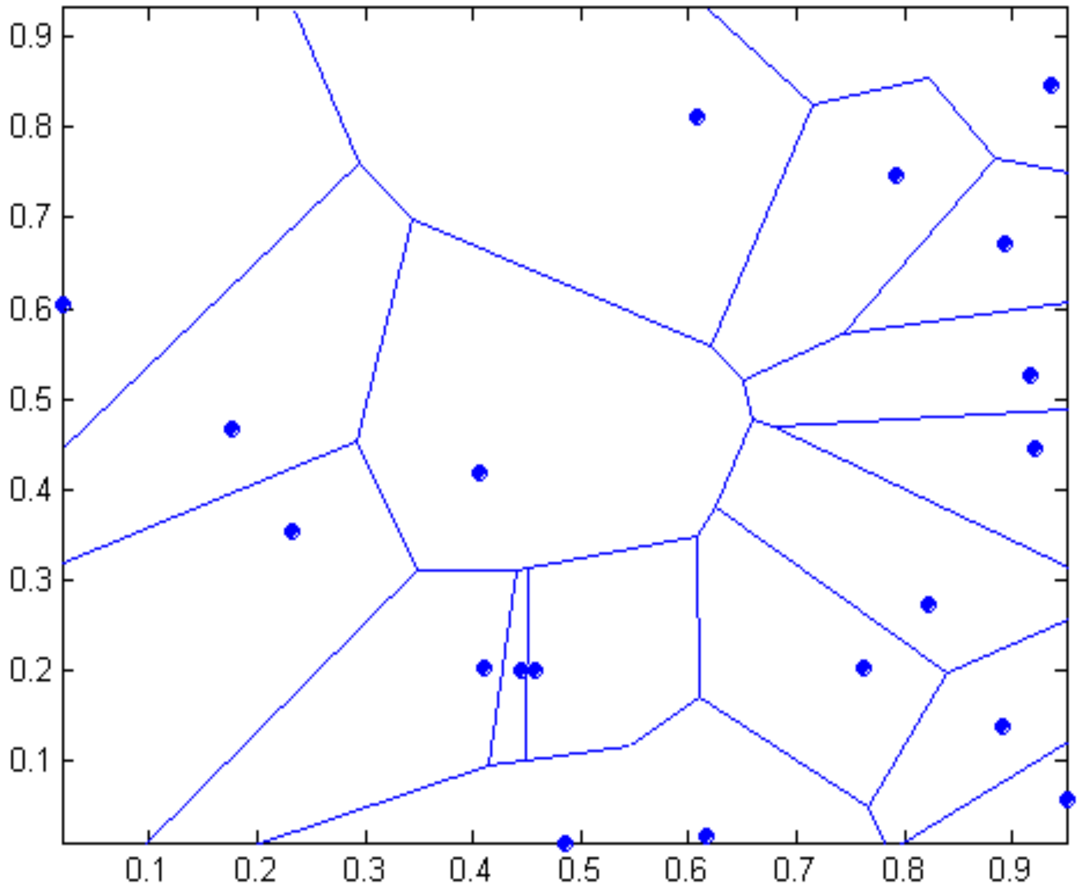


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distances to x

1 nearest-neighbor

Voronoi Diagram



Nearest Neighbor Classification

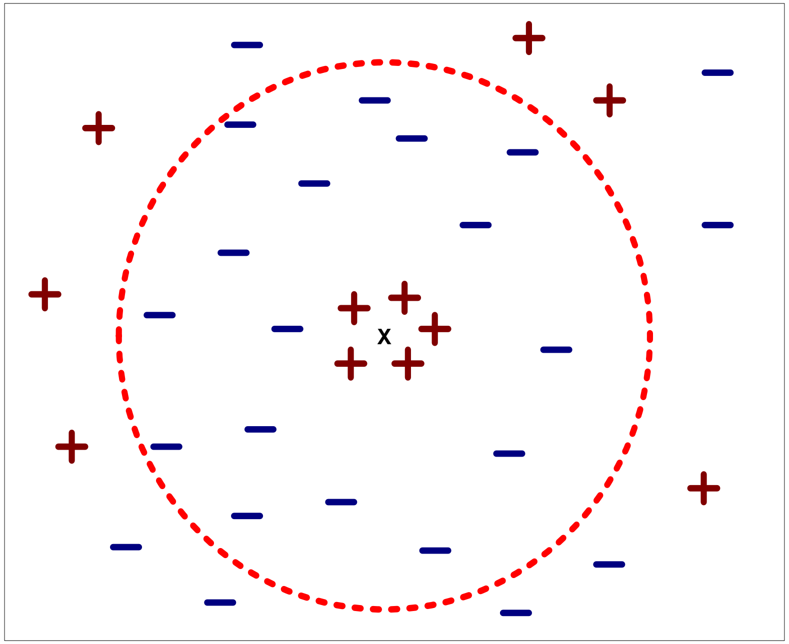
- Compute distance between two points:
  - Euclidean distance

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

- Determine the class from nearest neighbor list
  - Take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

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…

- Scaling issues
  - 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

Nearest Neighbor Classification…

- Selection of the right similarity measure is critical:

1 1 1 1 1 1 1 1 1 1 1 0		0 0 0 0 0 0 0 0 0 0 0 1
0 1 1 1 1 1 1 1 1 1 1 1	vs	1 0 0 0 0 0 0 0 0 0 0 0

Euclidean distance = 1.4142 for both pairs

Nearest neighbor Classification…

- k-NN classifiers are lazy learners since they do not build models explicitly
- Classifying unknown records are relatively expensive
- Can produce arbitrarily shaped decision boundaries
- Easy to handle variable interactions since the decisions are based on local information
- Selection of right proximity measure is essential
- Superfluous or redundant attributes can create problems
- Missing attributes are hard to handle

Improving KNN Efficiency

- Avoid having to compute distance to all objects in the training set
  - Multi-dimensional access methods (k-d trees)
  - Fast approximate similarity search
  - Locality Sensitive Hashing (LSH)
- Condensing
  - Determine a smaller set of objects that give the same performance
- Editing
  - Remove objects to improve efficiency

KNN and Proximity Graphs

- Proximity graphs
  - a graph in which two vertices are connected by an edge if and only if the vertices satisfy particular geometric requirements
  - nearest neighbor graphs,
  - minimum spanning trees
  - Delaunay triangulations
  - relative neighborhood graphs
  - Gabriel graphs
- See recent papers by Toussaint
  - G. T. Toussaint. Proximity graphs for nearest neighbor decision rules: recent progress. In Interface-2002, 34th Symposium on Computing and Statistics, ontreal, Canada, April 17–20 2002.
  - G. T. Toussaint. Open problems in geometric methods for instance based learning. In Discrete and Computational Geometry, volume 2866 of Lecture Notes in Computer Science, pages 273–283, December 6-9, 2003.
  - G. T. Toussaint. Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. Int. J. Comput. Geometry Appl., 15(2):101–150, 2005.