

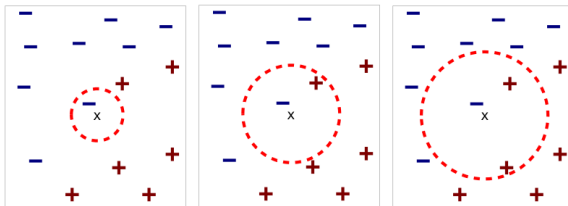
Start of the KNN: (K Nearest classification).

KNN does not require training.

Computing similarity.

All of the methods needs to calculate distance, requires normalization.

### 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 distance to x

### Nearest Neighbor Classification

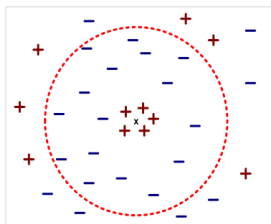
- Compute distance between two points:
  - Euclidean distance

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

- 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...

- 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