Lecture 02

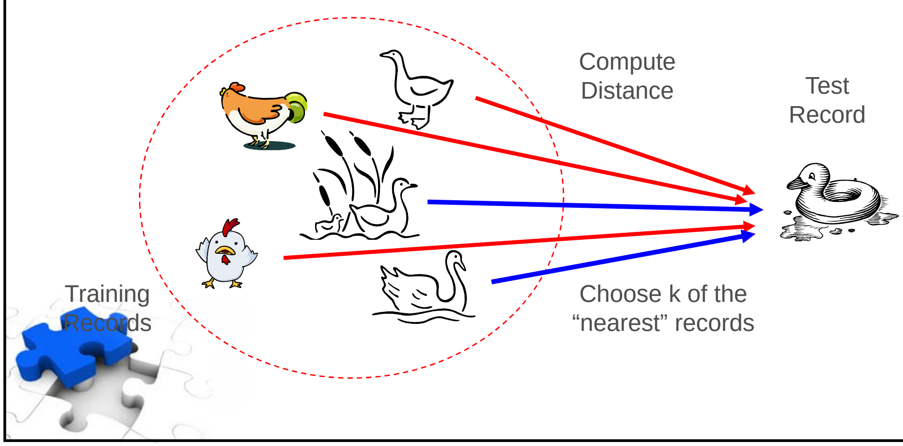
Classification: *k*-Nearest Neighbor Algorithm

Rawls Profess of MIS

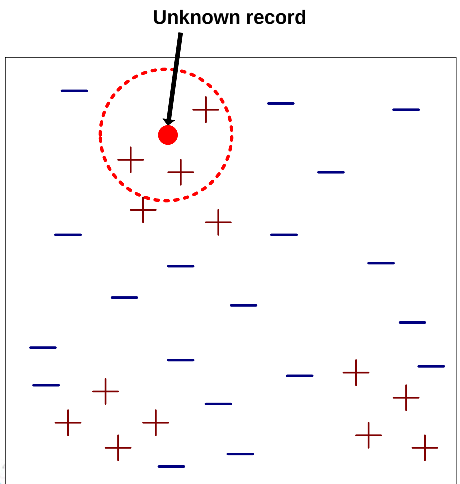
Jaeki Song, Ph.D.

Nearest Neighbor Classifiers

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



Nearest-Neighbor Classifiers

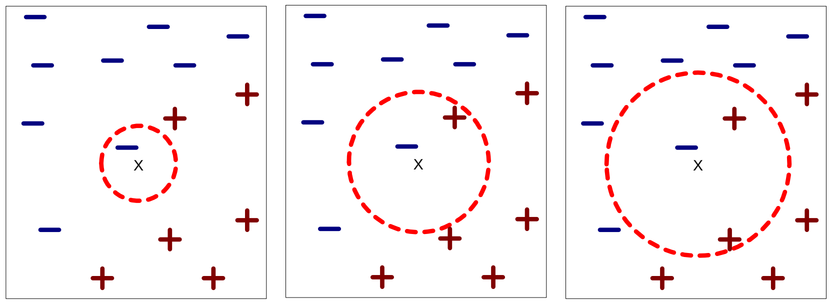


- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve
- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Issues..

- How many neighbors should we consider?
- How do we measure distance?
- How do we combine the information form more than one observation?
- Should all points be weighted equally?

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

- How is similarity defined between an unclassified record and its neighbors?
  - A distance metric is a real-valued function d used to measure the similarity between coordinates x, y, and z with properties:

1. $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(x,y)=\sqrt{\sum_i (x_i-y_i)^2}$$

  - Normalization
    - Continuous data values should be normalized using Min-Max Normalization or Z-Score Standardization

$$\text{Min-Max Normalization}=\frac{X-\min(X)}{\max(X)-\min(X)} \qquad \text{Z-Score Standardization}=\frac{X-\text{mean}(X)}{\text{standard deviation}(X)}$$

Nearest Neighbor Classification

- Which patient is more similar to a 50-year-old male: a 20-year-old male or a 50-year-old female?
  - For categorical attributes, the Euclidean Distance function is not appropriate
    - Instead, we define a function called “different”

$$\text{different}(x_i,y_i)=\begin{cases}0 & \text{if } x_i=y_i\\ 1 & \text{otherwise}\end{cases}$$

Nearest Neighbor Classification

- Let Patient A = 50-year-old male, Patient B = 20-year-old male, and Patient C = 50-year-old female
- Suppose that the Age variable has a range = 50, minimum = 10, mean = 45, and standard deviation = 15

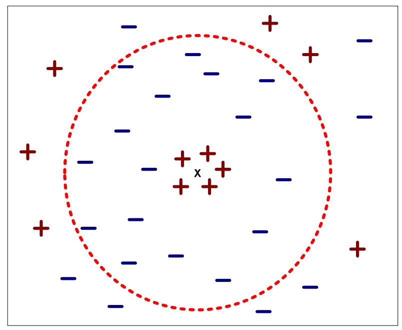
Nearest Neighbor Classification

- Different normalization techniques
  - resulted in Patient A being nearest to different patients in the training set
- The importance of understanding which technique is being used
  - Note that the distance(x,y) and Min-Max Normalization functions produce values in the range [0, 1]
- The distance between records containing both numeric and categorical attributes
  - Min-Max Normalization is preferred

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

How should the most similar (k) records combine to provide a classification?



Unweighted Voting

- This is the most simple combination function
- Decide on the value for k to determine the number of similar records that “vote”
- Compare each unclassified record to its k nearest (most similar) neighbors according to the Euclidean Distance function
- Each of the k similar records vote

Example

- 3-nearest neighbors classification

Instance	x1	x2	Class
1	5	7	1
2	4	3	2
3	7	8	2
4	8	6	2
5	3	6	1
6	2	5	1
7	9	6	2

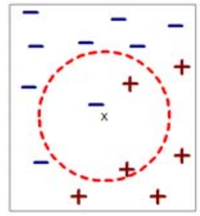
Example

	1	2	3	4	5	6	7
1	0.000	4.123	2.236	3.162	2.236	3.606	4.123
2	4.123	0.000	5.831	5.000	3.162	2.828	5.831
3	2.236	5.831	0.000	2.236	4.472	5.831	2.828
4	3.162	5.000	2.236	0.000	5.000	6.083	1.000
5	2.236	3.162	4.472	5.000	0.000	1.414	6.000
6	3.606	2.828	5.831	6.083	1.414	0.000	7.071
7	4.123	5.831	2.828	1.000	6.000	7.071	0.000

Instance	Nearest	Classification	Actual
1	3,4,5,	2	1
2	1,5,6,	1	2
3	1,4,7	2	2
4	1,3,7	2	2
5	1,2,6	1	1
6	1,2,5	1	1
7	1,3,4	2	2

Weighted Voting

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



- Ex

	1
1	0.000
2	4.123
3	2.236
4	3.162
5	2.236

$$record(3\&4) = \frac{1}{(2.236)^2} + \frac{1}{(3.162)^2} \cong 0.067$$

$$record(5) = \frac{1}{(2.236)^2} \cong 0.2$$

reverse

Estimation and Prediction

- The estimated target value $\hat{y}$ is calculated as

$$\hat{y}_{new} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

Choosing *k*

- Smaller k
  - Choosing a small value for k may lead the algorithm to overfit the data
  - Noise or outliers may unduly affect classification
- Larger k
  - Larger values will tend to smooth out idiosyncratic or obscure data values in the training set
  - It the values become too large, locally interesting values will be overlooked
- Choosing the appropriate value for k requires balancing these considerations

Exercise

- FileName: InClass02_LastName

Household	Income ($000s)	House Size	Ownership of Car
1	60	1840	Own
2	85.5	1680	Own
3	4.8	2160	Own
4	61.5	2080	Own
13	75	1960	lease
14	52.8	2080	lease
15	64.8	1720	lease
16	43.2	2040	lease

Source: Shmueli et al. (2016)