Data Mining

Lecture 11

Ananya Jana
CS360

Fall 2024

Instance Based Classifiers

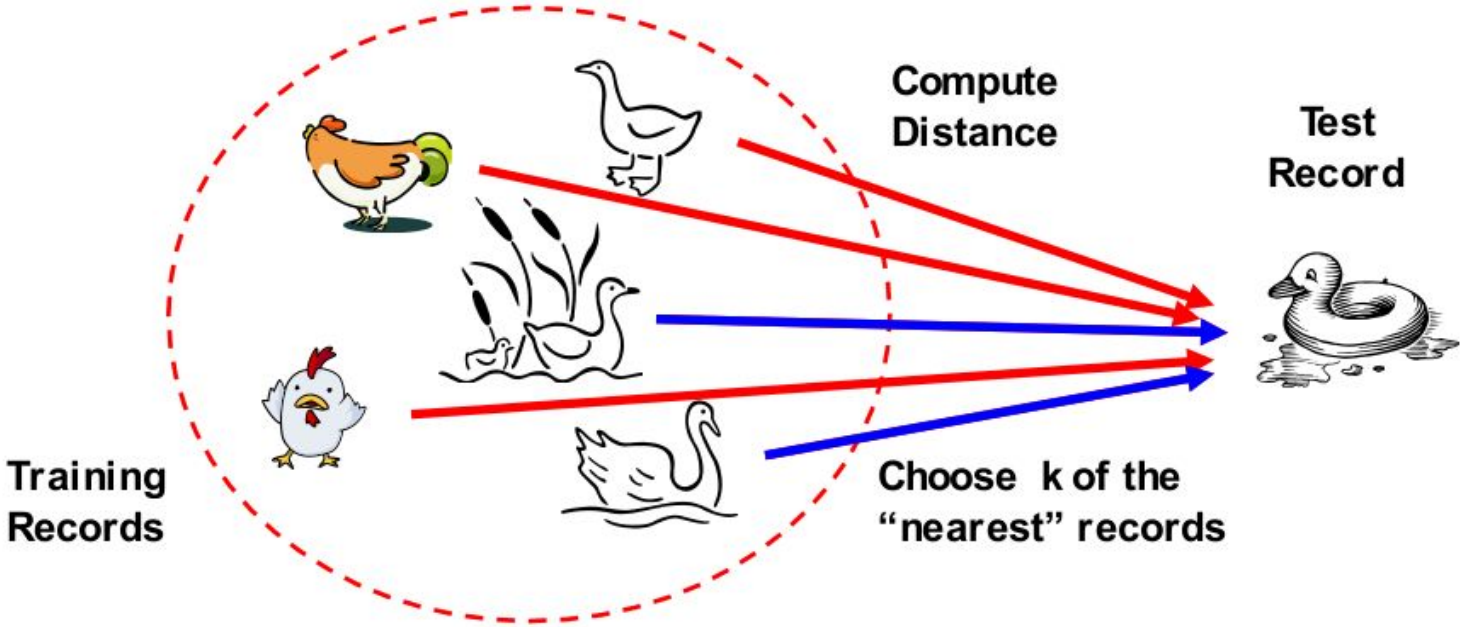
- Examples:
  - Rote-learner
    - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
  - Nearest neighbor
    - Uses k “closest” points (nearest neighbors) for performing classification

Instance Based Classifiers

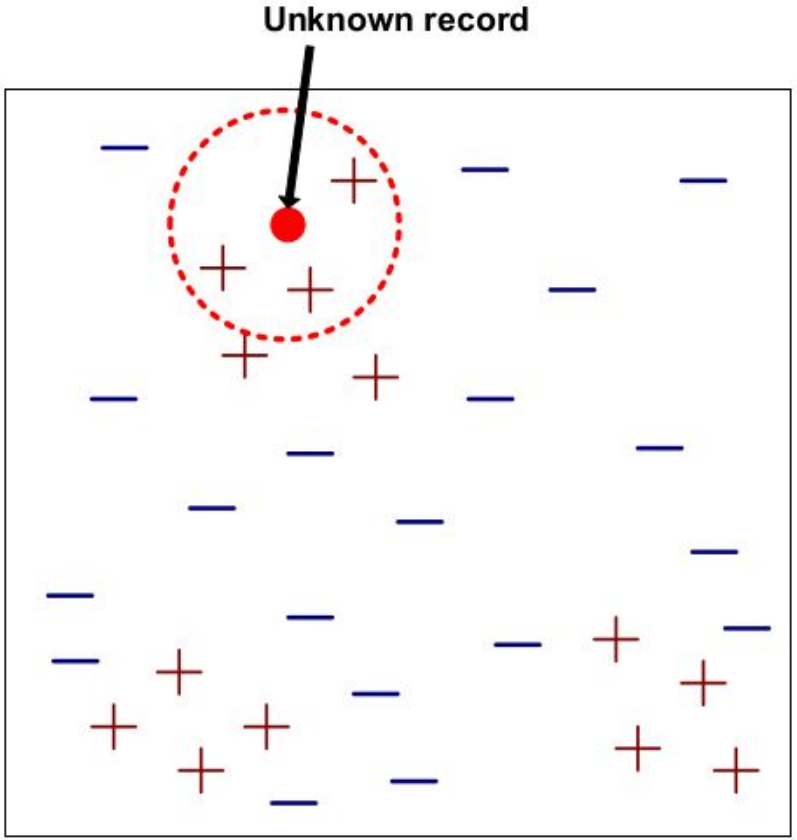
In K Nearest Neighbor algorithm, we find all the training examples that are relatively similar to the attributes of the test example. These examples, which are known as nearest neighbors, can be used to determine the class label of the test example

Nearest Neighbor Classifiers

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



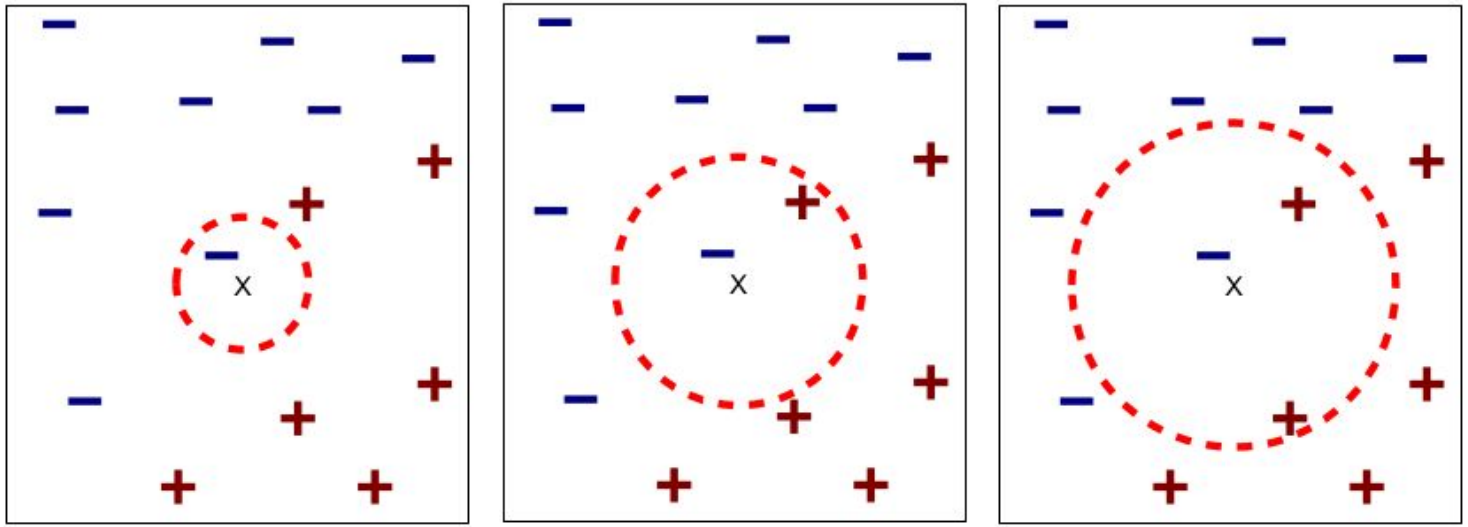
Nearest Neighbor Classifiers



- 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

Nearest Neighbor Classification

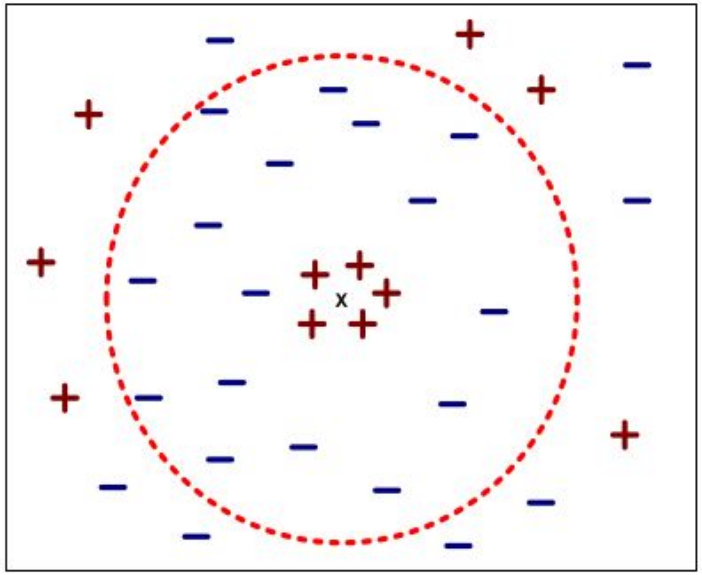
- Compute distance between two points:
  - Euclidean distance

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

- 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
	vs	
0 1 1 1 1 1 1 1 1 1 1 1		1 0 0 0 0 0 0 0 0 0 0 0

Euclidean distance = 1.4142 for both pairs

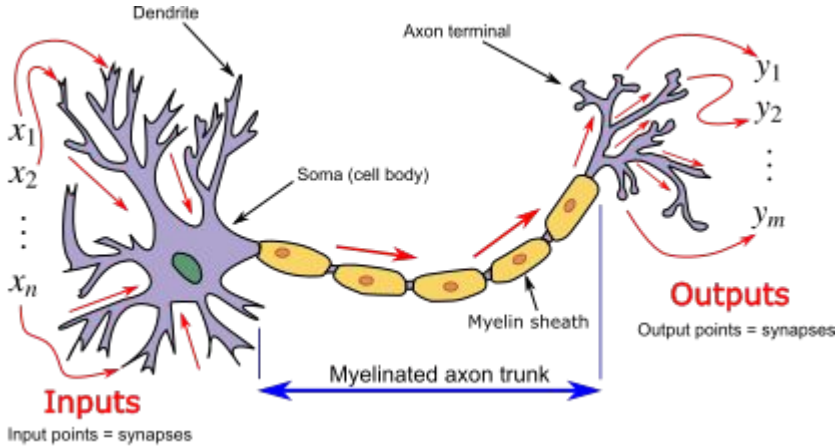
Artificial Neural Network

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.

Artificial neural networks (**ANNs**, also shortened to neural networks (NNs) or *neural nets*) are a branch of machine learning models that are built using principles of neuronal organization discovered by connectionism in the biological neural networks constituting animal brains.[1][2]

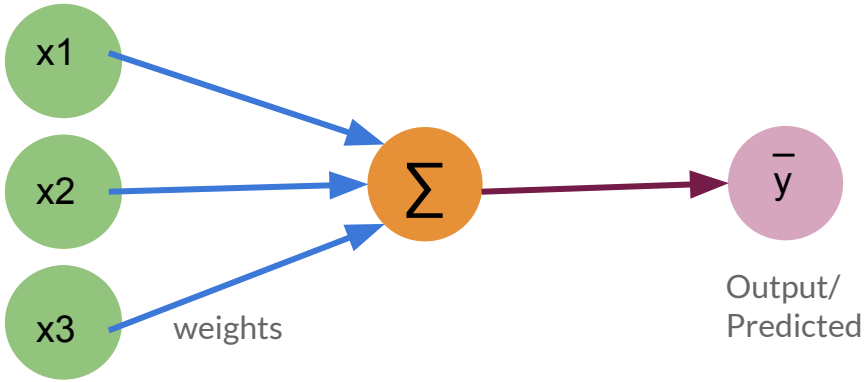
source:https://aws.amazon.com/what-is/neural-network/
Source: wiki

Artificial Neural Network



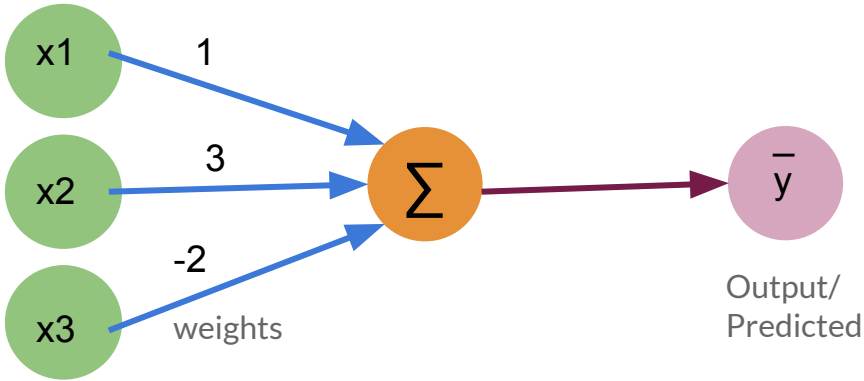
Neurons in human

Artificial Neural Network



x1	x2	x3	y(Actual)
5	6	7	10

Artificial Neural Network



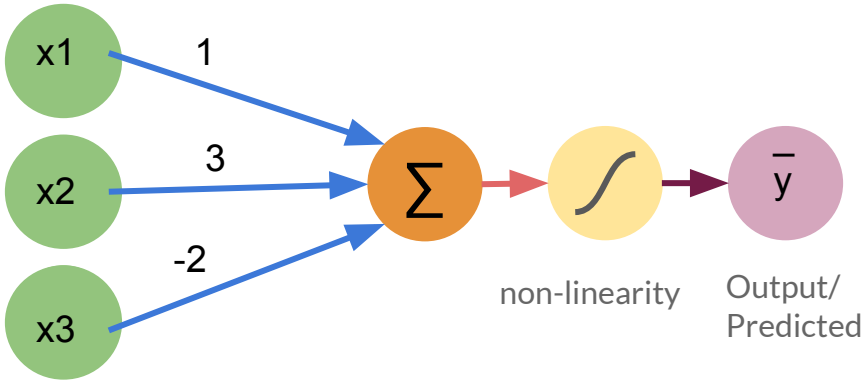
x1	x2	x3	y(Actual)
5	6	7	10

$\bar{y} = 1*5 + 3*6 + (-2)*7 = 9$

Error = predicted value - Actual Value
= 9 - 10
= -1

$\bar{y} = 1*x1 + 3*x2 + (-2)*x3$

Artificial Neural Network



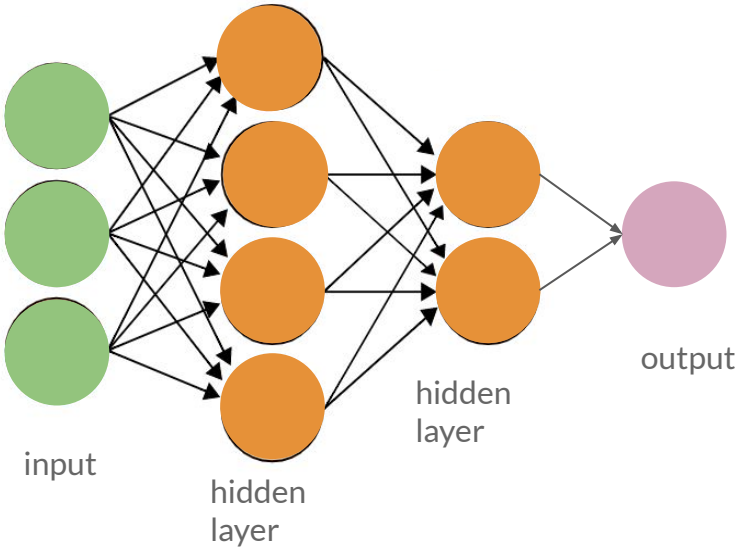
x1	x2	x3	y(Actual)
5	6	7	10

$\bar{y} = g(1*x1 + 3*x2 + (-2)*x3)$

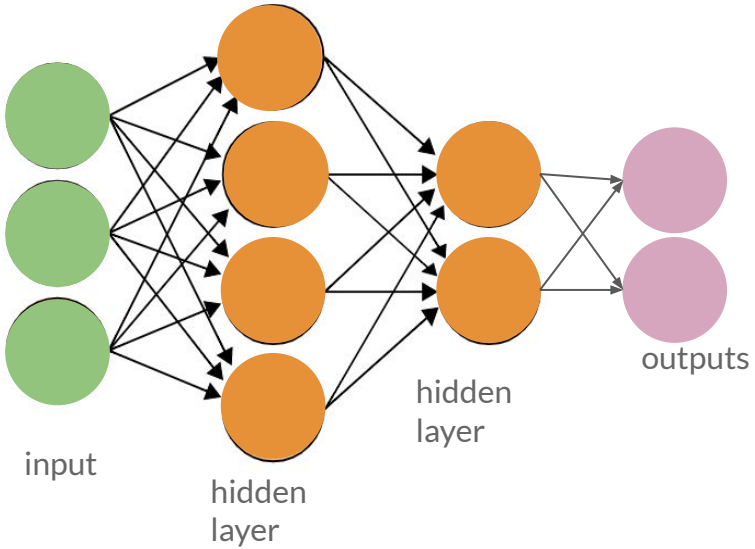
$\bar{y} = g(1*5 + 3*6 + (-2)*7) = g(9)$

Backpropagate the error

Artificial Neural Network



Artificial neural network with
multiple hidden layers



Artificial neural network with
multiple hidden layers and multiple outputs