Distance based methods in ML

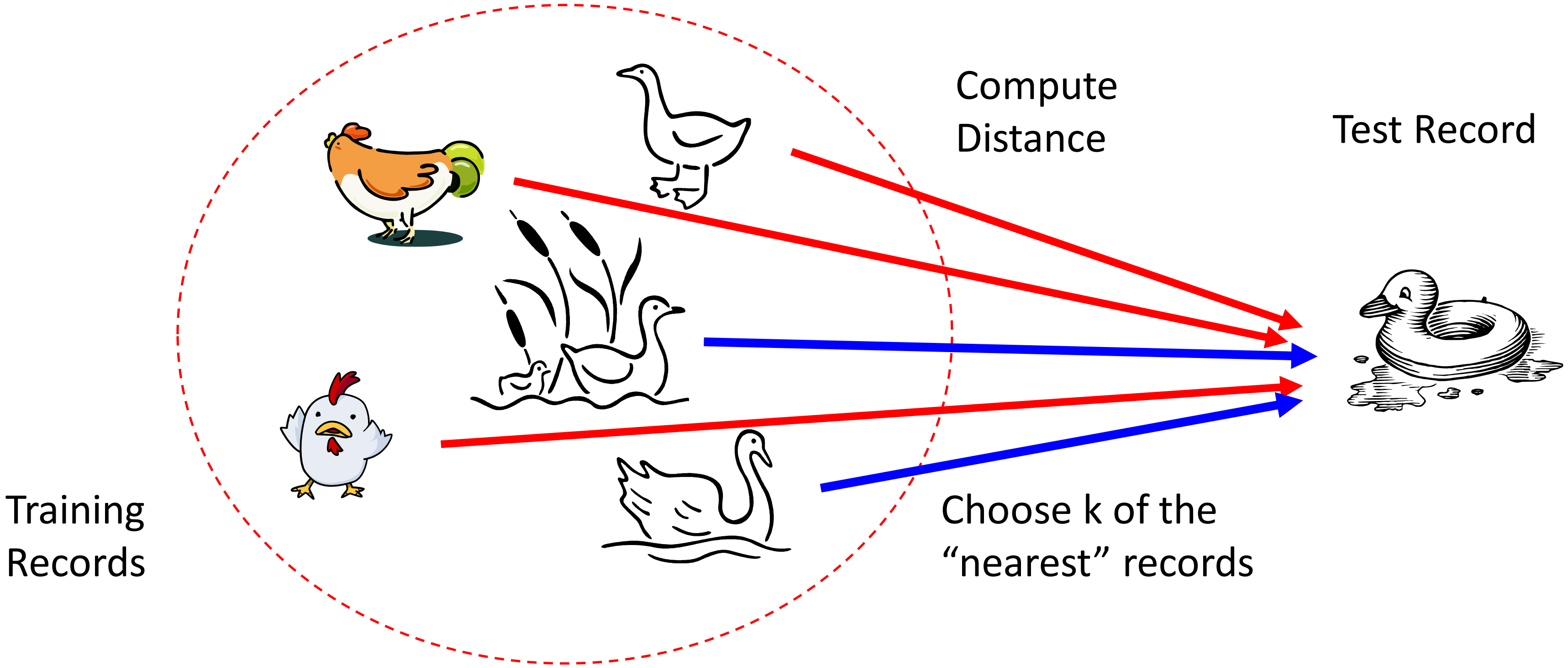
Distance measures play an important role in machine learning.

They provide the foundation for many popular and effective machine learning algorithms like k-nearest neighbors for supervised learning and k-means clustering for unsupervised learning.

A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain.

Nearest Neighbor Classifiers

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



Compute Distance

Test Record

Training Records

Choose k of the “nearest” records

Basic Idea of kNN

- *k*-NN classification rule is to assign to a test sample the majority category label of its *k* nearest training samples. Meaning given some training data and a new data point we would assign the new data based on the class of the training data it is nearest to.
- In other words, in KNN algorithm, a classification or regression prediction is made for new examples by calculating the distance between the new example (row) and all examples (rows) in the training dataset. The k examples in the training dataset with the smallest distance are then selected and a prediction is made by averaging the outcome.
- Simplest of all machine learning algorithms, can be used for both regression and classification.
- *k*-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- *k*-NN is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- In practice, *k* is usually chosen to be odd, so as to avoid ties.
- Choosing value of *k* is important. Should not be too less or too high!
- <u>Applications</u> – Video recommendation, Document classification etc.

How does K-NN work?

- The K-NN working can be explained on the basis of the below algorithm:
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

Advantages

- **No Training Period**- KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction and because of this it is very time efficient in term of improvising for a random modeling on the available data.
- **Easy Implementation**- KNN is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be calculated using distance formula such as- Euclidian or Manhattan
- As there is no training period thus new data can be added at any time since it wont affect the model.

Disadvantages

- **Does not work well with large dataset** as calculating distances between each data instance would be very costly. So, its computationally expensive.
- **Does not work well with high dimensionality** as this will complicate the distance calculating process to calculate distance for each dimension.
- **Sensitive to noisy and missing data -** Noise and mislabeled data, as well as outliers and overlaps between data regions of different classes, lead to less accurate classification.
- **Requires Good Choice of K**

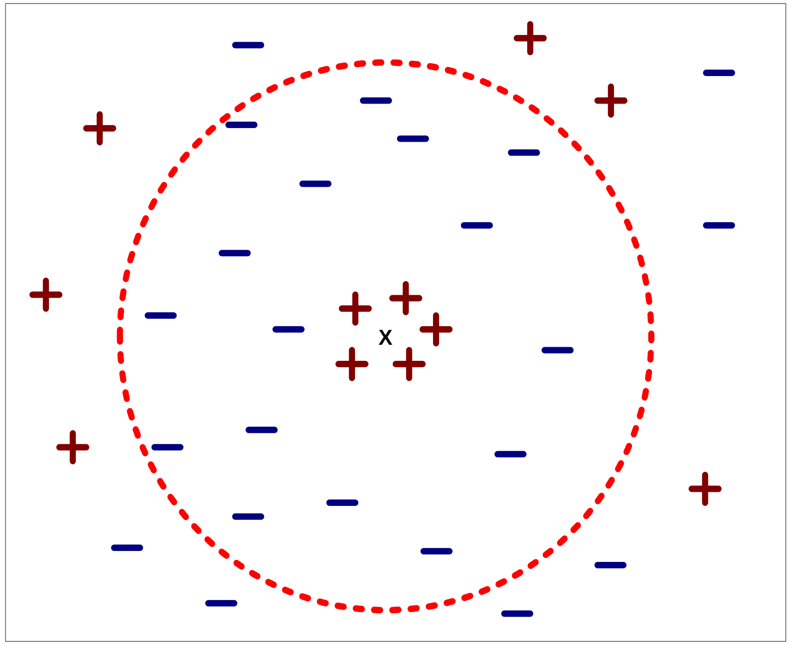
Nearest-Neighbor Classifiers: Issues

- The value of *k*, the number of nearest neighbors to retrieve
- Choice of Distance Metric to compute distance between records
- Computational complexity
  - Size of training set
  - Dimension of data

Value of K

- 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

Rule of thumb:
K = sqrt(N)
N: number of training points



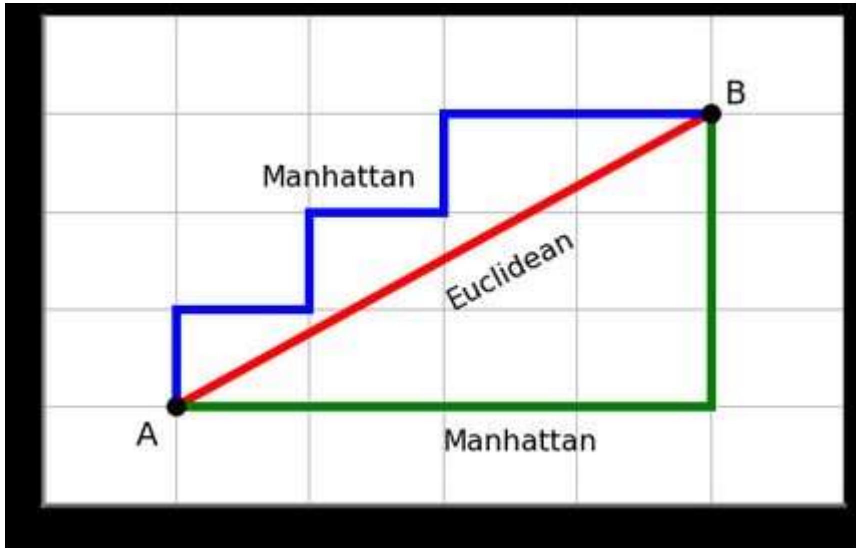
Distance Metric Widely Used :-

1.Euclidean Distance

$$d(x,\ y)\ =\sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

2.Manhattan Distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



3.Minkowski Distance

4.Hamming Distance

Thank You

Prepared by-

Mrs. Jheelam Mondal

Asst. Professor, CSE Department