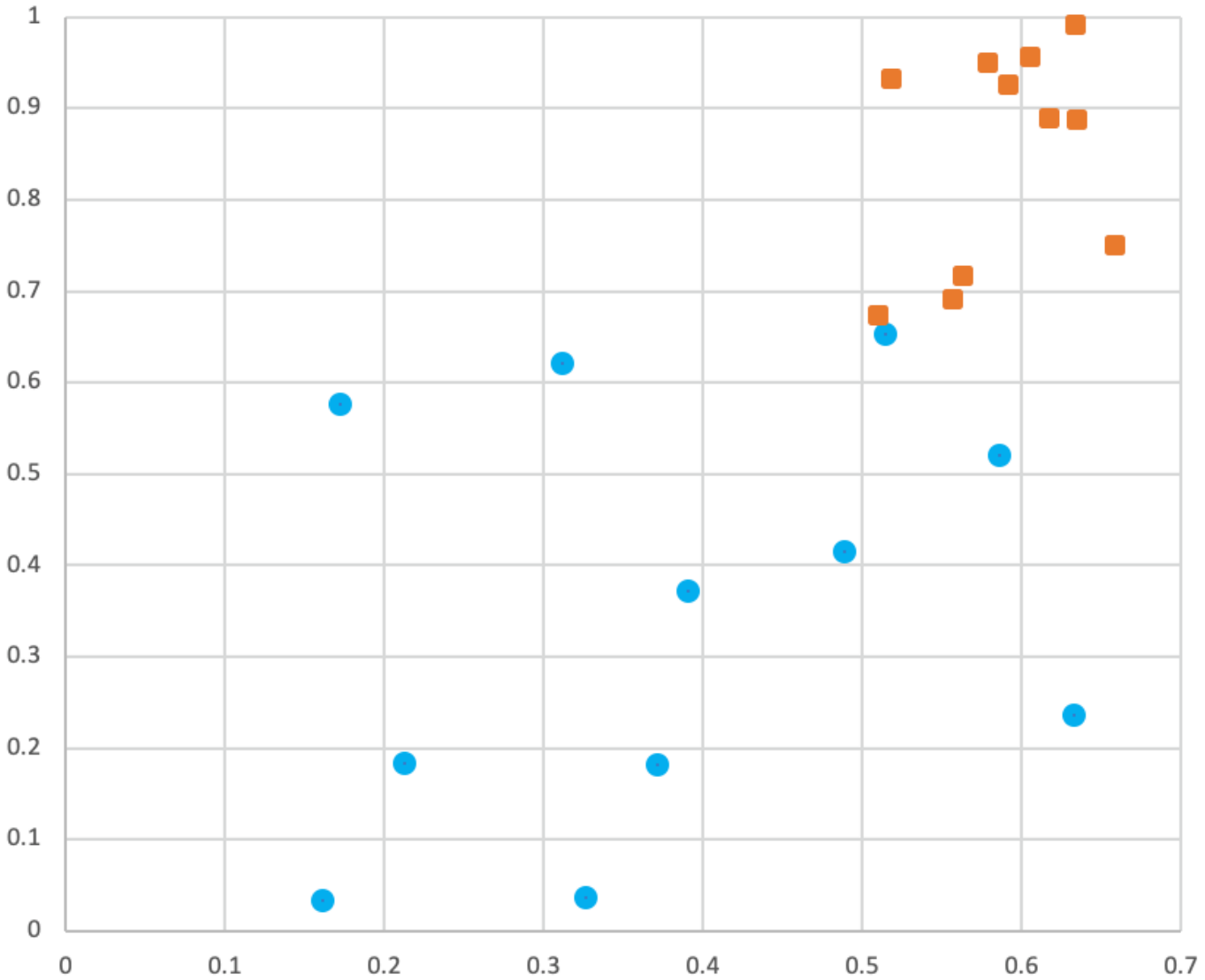
greatlearning
Learning for Life

Machine Learning

K-Nearest Neighbors

K-Nearest Neighbors

- Simple! A data point is most similar to its neighbors



Source: Wikipedia

Distance measure is important

- Most commonly distance is measured using Euclidian distances
- We should always Normalize data
- Other distance measurement methods include
  - Manhattan distance
  - Minkowski distance
  - Mahalanobis distance
  - Cosine similarity

- The approach to find nearest neighbors using distance between the query point and all other points is called the brute force. Becomes time costly and inefficient with increase in number of points
- Determining the optimal K is the challenge in K Nearest Neighbor classifiers.
    - Larger value of K suppresses impact of noise but prone to majority class dominating
-

Other Variants

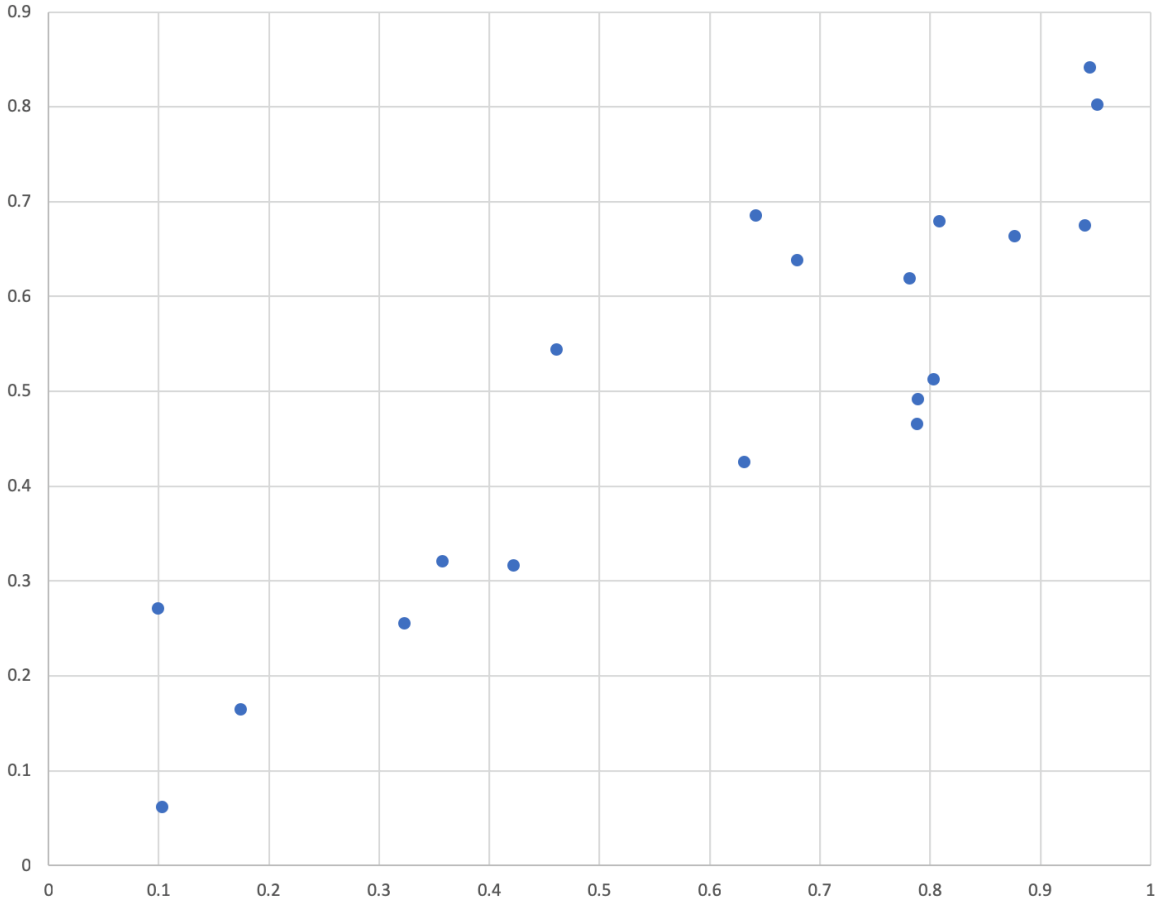
- Radius Neighbor Classifier
  - implements learning based on number of neighbors within a fixed radius r of each training point, where r is a floating point value specified by the user
  - may be a better choice when the sampling is not uniform. However, when there are many attributes and data is sparse, this method becomes ineffective due to curse of dimensionality
- KD Tree nearest neighbor
  - Approach helps reduce the computation time.
  - Very effective when we have large data points but still not too many dimensions

K-NN

- It does not construct a “model”. Known as a non-parametric method.
- Classification is computed from a simple majority vote of the nearest neighbors of each point
- Suited for classification where relationship between features and target classes is numerous, complex and difficult to understand and yet items in a class tend to be fairly homogenous on the values of attributes
- Not suitable if the data is too noisy and the target classes do not have clear demarcation in terms of attribute values
- Can also be used for regression

K-NN for regression

- The Neighbors based algorithm can also be used for regression where the labels are continuous data and the label of query point can be average of the labels of the neighbors



K Nearest Neighbors - pros and cons

- Advantages
  - Makes no assumptions about distributions of classes in feature space
  - Can work for multi classes simultaneously
  - Easy to implement and understand
  - Not impacted by outliers
- Dis-advantages
  - Fixing the optimal value of K is a challenge
  - Will not be effective when the class distributions overlap
  - Does not output any models. Calculates distances for every new point (lazy learner)
  - Computationally intensive