k - Nearest Neighbors

Classification Algorithm

kNN

- Given N training data points, kNN identifies k nearest neighbors of some data point c, regardless of labels

kNN

- Given N training data points, kNN identifies k nearest neighbors of some data point c, regardless of labels
- No probabilities involved!

kNN

- Given N training data points, kNN identifies k nearest neighbors of some data point c, regardless of labels
- No probabilities involved!
- Majority Vote!
    - Of all k nearest neighbors

kNN

- Given N training data points, kNN identifies k nearest neighbors of some data point c, regardless of labels
- No probabilities involved!
- Majority Vote!
    - Of all k nearest neighbors
- Classify based on distances from other data points

Minkowski Distance

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Minkowski Distance

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- Manhattan Distance
  - p = 1
- Euclidean Distance
  - p = 2

Cosine Distance

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos \theta$$

$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Cosine Distance

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

- Find similarity between 2 vectors
  - Cos 0 = 1 (similarity)
  - Cos 90 = 0 (some similarity)
  - Cos 180 = -1 (no similarity)

k-Nearest Neighbors (Algorithm)

- Determine the value of k
- For each testing data point
    - Find distance between the test point and all training points
    - Sort the distances (small to large) and select the top k testing points
    - Find the simple majority label from the k candidates
    - This is the predicted label for the test data point