Chapter 6 Similarity, Neighbors and Clusters

12/9/19 4:00 PM

Nearest Neighbor Reasoning

- Find k nearest observations by Euclidian distance
- Average the k known labels
  - For classification: simply by voting
  - For probability: weights of voting from a single label
  - For regression: (average) of the values of k neighbors
  - Need scaling before! Especially for Euclidean distance metric
- Weighted voting
  - Reduce influence on number of neighbors
  - Weighted by distance
- K is a complexity control parameter
  - 1-NN: strongly overfit (will predict the instance itself always, thus perfect fit)
  - Choosing K: model selection procedure with nested hold-out test, or nested CV, to see performance on training (validation) data
- Issues with K-NN
  - Intelligibility: hard to interpret what information mined from data set (features)
  - Curse of dimensionality
    - Fix 1: feature selection, to control number of features
    - Fix 2: manually adjust distance, e.g. assign more weights on important features, using domain knowledge
  - Computing efficiency is low for prediction, since the whole data set will be queried

Clustering

- Find natural groups in data set, also called unsupervised segmentation

1. *Hierarchical clustering*

a. Dendrogram: creates a collection of ways to group data
b. Linkage function: distance function to determine relationship between points

2. *K means clustering*

Procedure:

- Choose initial centroids (may be some selected data points)
- For every data point, calculate distance between data point to all centroids, and assign to the cluster with shortest distance
- Centroids would shift after all points are assigned
- Repeat for assigning every data point based on new centroids
- Until no change of cluster for all data points

Efficiency

- K-means generally higher efficiency, as only need to calculate points' distances to each centroids
- Hierarchical generally lower efficiency, need to calculate distance between each clusters, which initially is for every data point

Understand the result of clusters

- Names of each point in clusters can be useful if we need to target
- "Examplar" (best of class, for example), useful when clusters are large. Better than selecting random representatives in each cluster
- Centroid information (average characteristics, for example)