STAT 3690 Lecture 34

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 22, 2022

Clustering

- Problem: given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ group the observations into K populations
  - Unknown K
  - Unsupervised: no label/training data
- Why
  - Summarize a representation of the full data set
  - Exploration for structure of the data
  - Checking the validity of pre-existing group assignments
  - Assistance for prediction: sometimes clustering prior to prediction
- Clustering $C : \mathbb{Z}^+ \to \mathbb{Z}^+$
  - $C(i) = k$: assign $\boldsymbol{x}_i$ to group k

K-means

- Within-cluster scatter

$$W = \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i:C(i)=k}\sum_{j:C(j)=k}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \sum_{k=1}^{K}\sum_{i:C(i)=k}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k\|_2^2$$

  - $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$: the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$
  - $\bar{\boldsymbol{x}}_k = n_k^{-1}\sum_{i:C(i)=k}\boldsymbol{x}_i$
  - Smaller W is better
- To minimize the within-cluster scatter

$$\min_C W = \min_{C,\boldsymbol{c}_1,\ldots,\boldsymbol{c}_K}\sum_{k=1}^{K}\sum_{i:C(i)=k}\|\boldsymbol{x}_i - \boldsymbol{c}_k\|_2^2$$

- Implementation:
  1. Specify K and start with an initial guess for $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$, then repeat
     a. Labeling each point based the closest center: for each i, put X_i to the kth cluster such that $\boldsymbol{c}_k$ is closest to $\boldsymbol{x}_i$
     b. Replacing each center by the average of points in its cluster: for each k, take $\boldsymbol{c}_k = \bar{\boldsymbol{x}}_k$
  2. Terminate when W doesn't change
- Comments
  - Always converge

  - No guarantee to lead to the smallest W
  - Typically run K-means multiple times and pick up the result with the smallest W
    * Since the resulting clustering depends on the initial cluster centers

- Example (`iris`)
- An application to image compression/color quantization
  - Basic idea: compress images by reducing the color palette of an image to K colors

Figure 1: Image compression with K-means clustering (http://opencvpython.blogspot.com/2012/12/k-means-clustering-2-working-with-scipy.html)