CHE374 Week 1

Sanzhe Feng

September 19, 2022

Motivation of Unsupervised Learning

Usually, it is hard and expensive to collect labeled data required for supervised learning.

- Generative modeling
- Self-supervised learning
- Compression

In this course, we will be looking at three applications of unsupervised learning:

- Clustering: recommendation/search results, marktet segmentation, social network analysis
- Dimensionality reduction: get rid of redundancy
- Data visualization

Clustering with the k-means algorithm

k-mean algorithm is used to find the clusters of data.

First, randomly select two cluster centroids. Compute the distance between the each data to the centroids, then we assign each data point to the centroid that is closest to it.

Then we update the location of the centroids by computing the mean of all the points that have been assigned to each cluster.

Reset all data points to be unassigned. Repeat the distance calculation and the relocation of centroids.

k-mean algorithm

- input: k number of clusters; training set $\{x^1...x^n\}$
- random initialization: k cluser centroids $\mu_1...\mu_k$
- Repeat: for i in 1 to m

    $c^{(i)}$ = index of cluster centroid closest to x^i

for k in 1 to k

    μ_k = mean of points assigned to cluseter k

Distortion in the k-means algorithm

Optimization objective

Cost function: $J(c^1, ..., c^m, \mu_1, ..., \mu_u) = \frac{1}{m} \sum ||x^i - \mu_c||^2$

2 steps of k-mean algorithms:

1. Cluster assignment step: minimizing cost function wrt to c while fixing μ
2. Move centroids step: minimizing cost function wrt μ while fixing c

Initialization

How to pick the value of K:

Strategy is to pick random points as K in the input domain or K training sets as the K initial cluster centroids.

Local Optima

Problem that caused by initialization.