CS540 Introduction to Artificial Intelligence

Lecture 15

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 18, 2022

Midterm

Admin

Midterm Discussion

Admin

Unsupervised Learning

Motivation

- Supervised learning: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Unsupervised learning: $x_1, x_2, ..., x_n$.
- There are a few common tasks without labels.

1. Clustering: separate instances into groups.
2. Novelty (outlier) detection: find instances that are different.
3. Dimensionality reduction: represent each instance with a lower dimensional feature vector while maintaining key characteristics.

Unsupervised Learning Applications

Motivation

1. Google News
2. Google Photo
3. Image Segmentation
4. Text Processing

Hierarchical Clustering

Description

- Start with each instance as a cluster.
- Merge clusters that are closest to each other.
- Result in a binary tree with close clusters as children.

Hierarchical Clustering Diagram

Description

Single Linkage Distance

Definition

- Usually, the distance between two clusters is measured by the single-linkage distance.

$$d\left(C_k, C_{k'}\right) = \min\left\{d\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the shortest distance from any instance in one cluster to any instance in the other cluster.

Complete Linkage Distance

Definition

- Another measure is complete-linkage distance,

$$d\left(C_k, C_{k'}\right) = \max\left\{d\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the longest distance from any instance in one cluster to any instance in the other cluster.

Average Linkage Distance Diagram

Definition

- Another measure is average-linkage distance.

$$d\left(C_k, C_{k'}\right) = \frac{1}{\left|C_k\right|\left|C_{k'}\right|} \sum_{x_i \in C_k, x_{i'} \in C_{k'}} d\left(x_i, x_{i'}\right)$$

- It is the average distance from any instance in one cluster to any instance in the other cluster.

Hierarchical Clustering 1

Quiz

Hierarchical Clustering 2

Quiz

Hierarchical Clustering 3

Quiz

Hierarchical Clustering 4

Quiz

Hierarchical Clustering 4, Diagram

Quiz

Hierarchical Clustering 5

Quiz

Number of Clusters

Discussion

- K can be chosen using prior knowledge about X.
- The algorithm can stop merging as soon as all the between-cluster distances are larger than some fixed R.
- The binary tree generated in the process is often called dendrogram, or taxonomy, or a hierarchy of data points.
- An example of a dendrogram is the tree of life in biology.

K Means Clustering

Description

- This is not K Nearest Neighbor.
- Start with random cluster centers.
- Assign each point to its closest center.
- Update all cluster centers as the center of its points.

K Means Clustering Demo

Description

Distortion

Distortion

- Distortion for a point is the distance from the point to its cluster center.
- Total distortion is the sum of distortion for all points.

$$D_K = \sum_{i=1}^{n} d\left(x_i, c_{k^\star(x_i)}(x_i)\right)$$

$$k^\star(x) = \underset{k=1,2,\ldots K}{\operatorname{argmin}} \; d(x, c_k)$$

Objective Function Counterexample

Definition

Gradient Descent

Definition

- When d is the Euclidean distance. K Means algorithm is the gradient descent when distortion is the objective (cost) function.

$$\frac{\partial}{\partial c_k} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - c_k\|_2^2 = 0$$

$$\Rightarrow -2 \sum_{x \in C_k} (x - c_k) = 0$$

$$\Rightarrow c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

K Means Clustering 1

Quiz

K Means Clustering 2

Quiz

Total Distortion 2

Quiz

Number of Clusters

Discussion

- There are a few ways to pick the number of clusters K.

1. K can be chosen using prior knowledge about X.
2. K can be the one that minimizes distortion? No, when $K = n$, distortion $= 0$.
3. K can be the one that minimizes distortion $+$ regularizer.

$$K^{\star} = \underset{k}{\operatorname{argmin}} \left(D_k + \lambda \cdot m \cdot k \cdot \log n \right)$$

- λ is a fixed constant chosen arbitrarily.

Initial Clusters

Discussion

- There are a few ways to initialize the clusters.

1. K uniform random points in $\{x_i\}_{i=1}^{n}$.
2. 1 uniform random point in $\{x_i\}_{i=1}^{n}$ as $c_1^{(0)}$, then find the farthest point in $\{x_i\}_{i=1}^{n}$ from $c_1^{(0)}$ as $c_2^{(0)}$, and find the farthest point in $\{x_i\}_{i=1}^{n}$ from the closer of $c_1^{(0)}$ and $c_2^{(0)}$ as $c_3^{(0)}$, and repeat this K times.

Gaussian Mixture Model

Discussion

- In K means, each instance belong to one cluster with certainty.
- One continuous version is called the Gaussian mixture model: each instance belongs to one of the clusters with a positive probability.
- The model can be trained using Expectation Maximization Algorithm (EM Algorithm).

Gaussian Mixture Model Demo

Discussion