Clustering

Sunday, October 28, 2018 2:52 PM

2016:

1. **[Clustering, K MEANS:]**

The k-means algorithm iteratively computes the set of centroids given a clustering of the data points, and the clustering of the data points given a set of centroids. In this question, you will provide formulas for the iterative steps of the k-means algorithm.

Let the data points be $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in \mathbb{R}^d$.
Let the clusters be subsets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k \in \{1, 2, \ldots, n\}$ of the indices.
Let the centroids be d-dimensional vectors $z^{(1)}, \ldots, z^{(k)} \in \mathbb{R}^d$.

a. Suppose that you are given the clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k \in \{1, 2, \ldots, n\}$. Write down the formula for each of the centroids $z^{(1)}, \ldots, z^{(k)} \in \mathbb{R}^d$.

$$z^{(j)} = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x^{(i)}$$

b. Suppose that you are given the centroids $z^{(1)}, \ldots, z^{(k)} \in \mathbb{R}^d$. Write down the quantity that we need to minimize to find the cluster $\mathcal{C}_j$ for a particular data point $x^{(i)}$.

$$\left\| x^{(i)} - z^{(j)} \right\|$$

c. The cost function in the k-means algorithm is not convex, so it could have local minima that give rise to poor clustering. Briefly describe one strategy for overcoming this issue.

Ans: We could try many **random initializations** of the **centroids**, and **run the k-means algorithm for each initialization.** We then **pick** the **clustering** that **minimizes** the **cost** of clustering.

$$\text{cost}(\mathcal{C}, z) = \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| x^{(i)} - z^{(j)} \right\|^2$$

Alternatively, we can initialize the centroids far apart from each other, using k-means++.

d. To find the optimal number k of clusters, a method called validation is often used. Describe the steps involved in validation. In particular, state the performance metric used for computing the validation error in k-means clustering.

Ans: First, the available data is partitioned into a **training set** and a **validation set** (usual split is around 70% and 30%). For each k, the **k-means algorithm** is performed several times on the training set using different initializations and the best result is picked. The **cost of clustering is computed on the validation set.** Using this cost as the validation error, we **plot a graph of the validation error against the number of clusters.** The **'elbow'** point after which the cost of clustering does not change much is **picked as the optimal number of clusters.**

2017:

1. CLUSTERING
2. **[Clustering | Calculating Centroid and boundary of Voronoi Regions]**

The k-means algorithm iteratively computes the set of centroids given a clustering of the data points, and the clustering of the data points given a set of centroids. In this question, you will analyze the performance of the k-means algorithm on a one-dimensional problem. Suppose that we have six data points $1,2,3,50,100,150 \in \mathbb{R}$.

a. **If k=1 in your k-means algorithm, where would the centroid of the single cluster be?**
   Ans: centroid = (1+2+3+50+100+150)/6
   = 51

b. **If k=2 in your k-means algorithm, there will be two Voronoi regions corresponding to the two clusters. The boundary between the Voronoi regions will be:**
   Ans: Exactly halfway between the two centroids.

c. If k=2 in your k-means algorithm, which of the following are possible clusters when the algorithm has converged? Circle 'Yes' if it is a possible clustering, and 'No' otherwise.

   Ans:

   - {1, 2} and {3, 50, 100, 150} Yes / No
   - {1, 2, 3} and {50, 100, 150} Yes / No
   - {1, 2, 3, 50} and {100, 150} Yes / No
   - {1, 2, 3, 50, 100} and {150} Yes / No