k-Means
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

19 Jul 2022

1. Introduction

k-means is an unsupervised machine learning algorithm that is used for classification. It achieves this by clustering the input data set into k clusters based on the proximity of points with other. The basis of this algorithm is the assumption that points closer to each other are similar and points far from each other are not.

k points are chosen randomly and are assigned k different labels. All the other points are assigned labels based on the 1-nearest neighbour classifier. This leads to the formation of k clusters. Then, the centroid of each cluster is computed and all the points are re-assigned labels based on their distances from the cluster centroids. This process is repeated until the assignment of labels becomes stable, i.e., there is no change in the assignment from the current state to the next state.

Since the intial points are chosen randomly, this might not be the best classifier. Keep repeating the process by choosing a different set of k random points each time until the total variance in the clusters is minimized.

2. Mathematical Formulation

Suppose we are given m samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ where $x_i \in \mathbb{R}^n \ \forall i \in [m]$
Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be a metric that denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$
The algorithm is as follows:

i) Choose k random points $\mathbf{m}_1^{(0)}, \mathbf{m}_2^{(0)}, \ldots, \mathbf{m}_k^{(0)}$ to be our initial cluster centres.

ii) Assign the label $y_i \in [k]$ to each $\mathbf{x}_i$ based on the nearest cluster centre at iteration t

$$y_i^{(t)} = \hat{j} \text{ where } \hat{j} = \arg\min_{j \in [k]} d(\mathbf{x}_i, \mathbf{m}_j) \tag{2.1}$$

iii) Set the centroids of the obtained clusters as the new cluster centres.

$$\mathbf{m}_j^{(t+1)} = \frac{1}{\left|S_j^{(t)}\right|} \sum_{\mathbf{x}_i \in S_j^{(t)}} \mathbf{x}_i \text{ where } S_j^{(t)} = \{\mathbf{x}_i \mid y_i = j\} \tag{2.2}$$

The performance of the model is quantified using the total variance in the clusters. The goal of the k-means algorithm is to find a clustering $\mathbf{S} = \begin{pmatrix} S_1 & S_2 & \ldots & S_k \end{pmatrix}^\top$ that will minimize the total variance.

$$\hat{\mathbf{S}} = \arg\min_{\mathbf{S}} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in S_j} d(\mathbf{x}_i, \mathbf{m}_j) \tag{2.3}$$

3. Choosing the Optimal k

Choosing the right k is important. As the value of k increases, the total variance in clusters also decreases. However, if k is very large, the entire purpose of clustering itself is defeated and it is also computationally expensive. This is why we look for a trade-off.

When the reduction in variance is plotted against k, we obtain an elbow plot. There is a point at which there is a sharp decrease in the reduction in variance, which is known as the elbow. Beyond this point, the diminishing returns are no longer worth the additional cost. The elbow is chosen as the value of k.

4. Questions

i) How is k-means different from k-nearest neighbours?
ii) Is the algorithm guaranteed to converge?
iii) Is the algorithm guaranteed to find the optimal clustering?
iv) Give a practical application of k-means clustering.
v) What is a drawback of k-means?

5. Answers

i) k-means is an unsupervised machine learning algorithm that determines the categories of the input data whereas k-nearest neighbours is a supervised machine learning algorithm that classifies a data point into one of the given categories.
ii) When the metric d is chosen to be the square of the Euclidean distance, then the algorithm will converge. But in general, it might not converge.
iii) The algorithm is not guaranteed to find the optimal clustering because the inital cluster centres are random. This is why the algorithm is repeated multiple times with different initial cluster centres to find the optimal clustering.
iv) k-means is often used in image segmentation. It is used for colour quantization which is the reducing of the colour palette of an image to k colours.
v) Noisy input can drastically hamper the performance of a k-means model because they affect the calculation of the cluster centroids. A noisy data point can cause the cluster centroid to be much farther away than it is supposed to be.