K means and simple learning-Lecture 1

28 January 2024 11:58

mmml_mlt1

mmml_mlt1

Reference List :
https://viso.ai/deep-learning/supervised-vs-unsupervised-learning/
Accessed on 2nd Feb

Goals of Cluster analysis:

Hastie, Tibshirani & Friedman:
The Elements of Statistical Learning
https://web.stanford.edu/~hastie/ElemStatLearn/

What is K-means Algorithm?

The K-means algorithm is a popular clustering technique used to partition a dataset into K clusters based on similarity. Here's a breakdown of the key components mentioned:

1. Quantitative Variables: This refers to variables in the dataset that represent numerical quantities. In other words, the data consists of numerical values that can be measured and compared. For example, if you're clustering data related to customer demographics, quantitative variables might include age, income, number of purchases, etc.
2. Squared Euclidean Distance: The squared Euclidean distance is a measure of the dissimilarity between two points in a multidimensional space. It's computed as the sum of the squared differences between corresponding coordinates of the two points. For two points ppand qqin n-dimensional space, the squared Euclidean distance is calculated as:
$\sum_{i=1}^{n}(p_i - q_i)^2$

K-means relies on the concept of distance to group data points into clusters. It minimizes the sum of squared Euclidean distances between data points and their respective cluster centroids.

3. Iterative Descent: The K-means algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids based on the mean of the data points assigned to each cluster. This process continues until the cluster assignments and centroids converge or until a predefined stopping criterion is met (e.g., a maximum number of iterations).
In summary, the K-means algorithm is suitable for datasets where all variables are numerical (quantitative) and the similarity between data points is measured using the squared Euclidean distance. It's a popular method for partitioning data into clusters based on these numerical attributes and is widely used in various fields such as machine learning, data mining, and pattern recognition.

From <https://chat.openai.com/c/c0126b7d-c773-4c76-936f-448f97da5d3b>

4. Distortion Function:
   - The distortion function, also known as the cost function or inertia, measures the average squared distance between each data point and its assigned cluster centroid in k-means clustering.
   - Mathematically, for a cluster with centroid □□and a set of data points □□, the distortion for that cluster is calculated as the sum of squared distances:
   - The overall distortion for the entire clustering solution is the sum of distortions across all clusters.
5. Within-Cluster Sum of Squares (WCSS):
   - WCSS is a term commonly used interchangeably with distortion in the context of k-means clustering.
   - It is essentially the same as the distortion function and represents the sum of squared distances between each data point and its assigned cluster centroid across all clusters.
   WCSS=∑□=1□∑□∈□□□□2WCSS=∑i=1k∑x∈Xi∥x−ci∥2
   - Here, □□is the number of clusters, □□□is the set of data points in the i-th cluster, and □□is the centroid of the i-th cluster.

From <https://chat.openai.com/c/0ef642e8-c2c6-45bd-a0ef-d426ea6349c9>

Unsupervised ML:

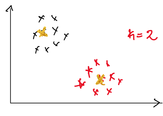
We dont have any specific o/p, just the data points.

where is it used?

It can be used before Ensembling techniques

K Means:

K → No: of centroids



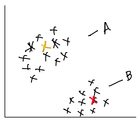
We try different values for K

K= 1, 2, 3. . .

Once we select say K=2, lets randomly select any 2 data points & initialise them as centroids.

Now we find the distance of the points from each of the centroid & is allocated to the corresponding group is assigned.

step 1



Note that centroid position will be updated by finding the mean of all the points of that group.

The distance of the data points are calculated from each of the centroid & they are allocated to the centroid closest to them. This continues until the centroid doesnt get updated anymore

How to decide the K-value:

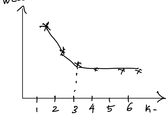
Elbow method:

We perform an iteration of K value for K= 1 – 10 & plot the graph btwn K & WCSS (within cluster sum of squares)

for K=1
WCSS = data point to cluster distance → will be very high

K=2

as K ↑, WCSS ↓.
(in this case)



an abrupt change of value point to the ideal no: of K.