CSC380: Principles of Data Science

Class Notes | Clustering - K Means

Table Of Contents

Supervised versus Unsupervised Learning

What is the Key Difference Between Unsupervised Learning versus Supervised Learning?

Supervised Learning: Learning under supervision, We have a full set of labeled data for learning.

Unsupervised Learning: They analyze and cluster unlabeled data sets using algorithms that discover hidden patterns in data without the need for human intervention, Unlabeled Data.

Read more at Nvidia Blog, IBM Blog

How can we view the task of discrete binning/grouping in unsupervised and unsupervised learning setups?

Supervised Learning: Classification
Unsupervised Learning: Clustering

Clustering

What is a Cluster?

A Cluster can be thought of as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster.

What is Clustering?

The grouping of objects such that objects in the same cluster are more similar to each other than they are to objects in another cluster.[1]

Or

The task of finding an assignment of data points to clusters, as well as a set of vectors {μk}, such that the sum of the squares of the distances of each data point to its closest vector μk (or another similarity measure) is a minimum.

Read more at (1) Nvidia Blog Google Developers Blog

What are the Different Types of Clustering?

Clean and simple explanation with diagrams at Google Developers Machine Learning Course - Clustering Algorithms

K Means

Based on Chapter 9: Pattern Recognition and Machine Learning- Christopher M Bishop

What is the Basic Idea behind K-means?

- Assign (Random) Cluster Centroids
- Until Convergence:
  - Cluster Assignment Step
  - Re-assigning Centroid Step

A simple gif visualization at Introduction to K-Means Clustering in Python with scikit-learn

Video Recommendation: Andrew NG - Machine Learning Course - This Lecture

What is the 1-of-K Coding Scheme?

For each data point xn, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{ 0, 1\}$, where k = 1, . . . , K describing which of the K clusters the data point x_n is assigned to, so that if data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for j = k.

What is the Objective Function in K-Means Algorithm?

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Also called the Distortion Measure, this represents the sum of the squares of the distances of each data point to its assigned vector μ_k. N is range of Data Points, and K is range of Topics/Classes.

Our goal is to find values for the $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J.

What is Distortion Measure?

Same as above question

What is the K-Means algorithm? (More Formalised Version)

1. Choose the number of clusters k
2. Select k random points from the data as centroids (We later discuss how to optimise this)
3. Until Convergence
    a. Assignment to a cluster Head/Centroid.

    $$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

    b. Resetting the Centroid.
    Find the mean of all points in a cluster, and set that as the new centroid of the cluster. (Hence the name k-means)

    $$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$$

Convergence in K Means

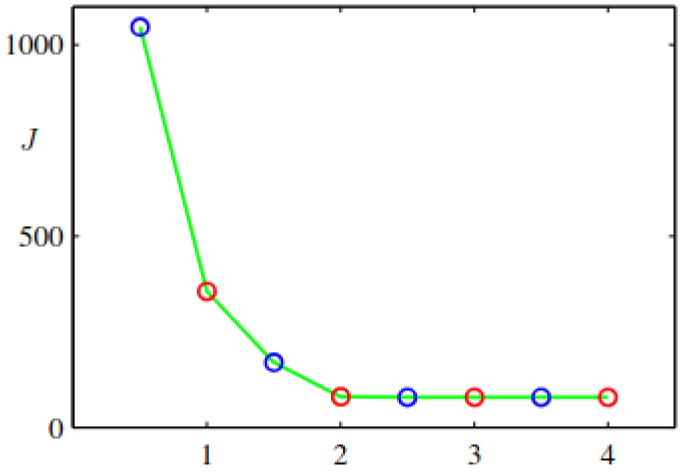
What is Convergence in K-Means?

There is no further change in the assignments of the centroid of clusters, or cluster assignment of points in the algorithm.

Is Convergence Guaranteed in K-Means?

Yes, in each phase our Objective Function will decrease and will reach a steady state.

An example plot:



Source: Fig 9.2 in Bishop - Pattern Recognition And Machine Learning

Minima Issues in K-Means

K-means may converge at a Local minima than a Global Minima.

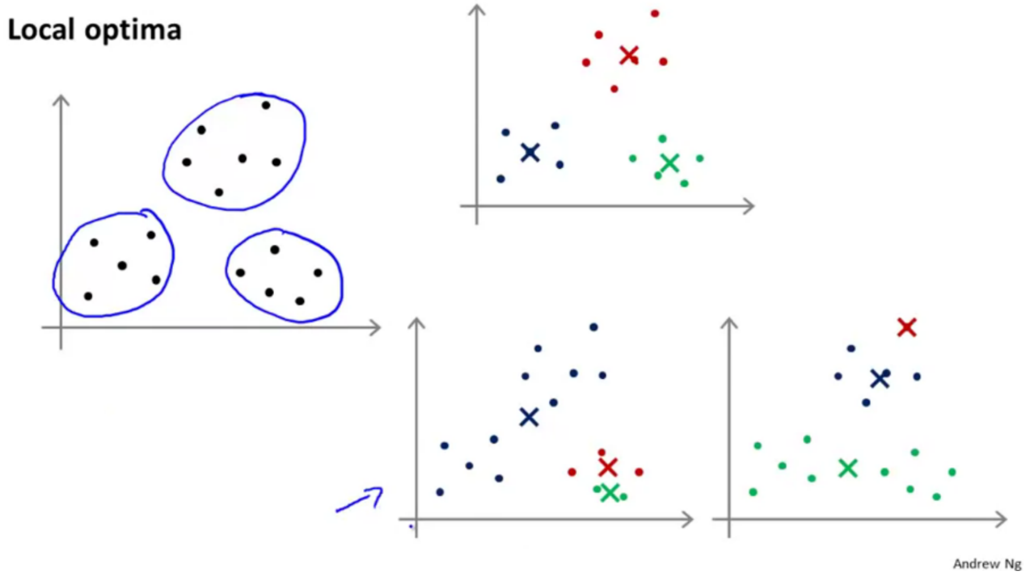


Fig Source: Andrew NG Coursera Machine Learning Course

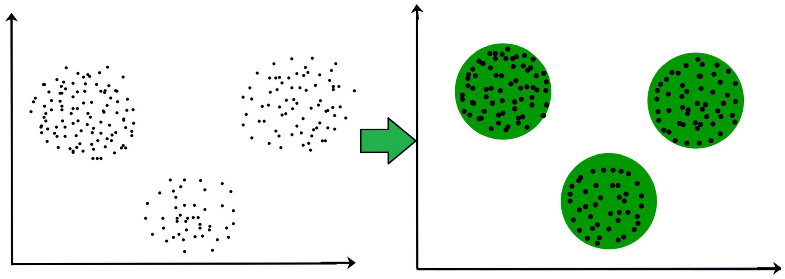
What are some reasons why a cluster may have just One Point?

1. The Cluster Point is An Outlier
2. Local Minima Was Attained.

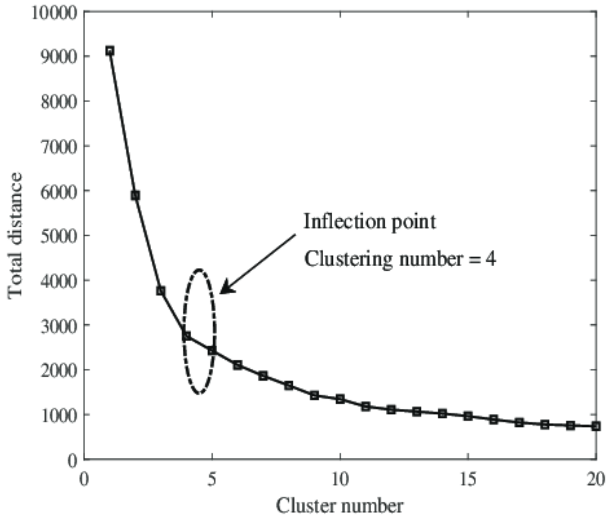
The Number Of Topics for K Means

How can we choose a number of Topics for K-Means?

1.The Most Common Approach is to Visualise it and Choose.



2. The Elbow Method



Over a range of k, we compute distortion score (or a similar metric). When these are plotted as a line chart, we will observe a point of inflection- an elbow.

This is a recommended number of Topics.

Source of Image

But sometimes, we may not observe a strict inflection point, like in the fig below.

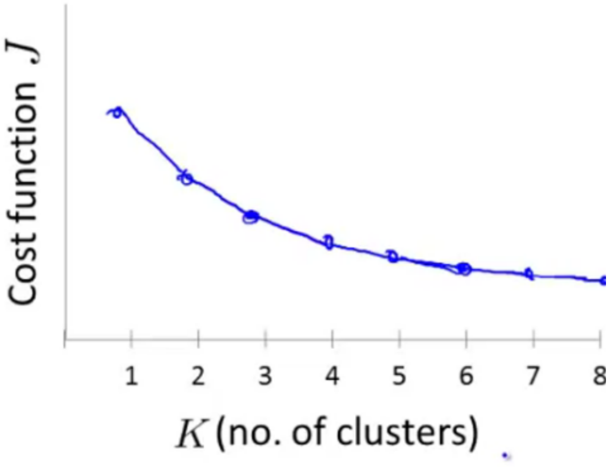


Image Source : Andrew NG Machine Learning Machine Learning Course

3. KMeans++ Algorithm

What is the KMeans ++ Algorithm?

KMeans does the first step, assignment of the first round of initialization of centroids in a smarter initialization of the centroids and improves the quality of the clustering.

From GeeksForGeeks:

1. Randomly select the first centroid from the data points.
2. For each data point compute its distance from the nearest, previously chosen centroid.
3. Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid. (i.e. the point having maximum distance from the nearest centroid is most likely to be selected next as a centroid)
4. Repeat steps 2 and 3 until k centroids have been sampled

Video Recommendation: Sara Jensen: K Means++

K Medoids

K-means may not be the most suitable algorithm in some cases, since it is very sensitive to noise and outliers. While, K-means attempts to minimize the total squared error, while k-medoids minimize the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids choose datapoints as centers (medoids or exemplars).[1]

So an improvised K means algorithm - K medoids is used.

K-medoids algorithm from GeeksForGeeks

```
1. Initialize: select $k$ random points out of the $n$ data points as the medoids.
2. Associate each data point to the closest medoid by using any common distance metric methods.
3. While the cost decreases:
    For each medoid m, for each data o point which is not a medoid:
        1. Swap m and o, associate each data point to the closest medoid, recompute the cost.
        2. If the total cost is more than that in the previous step, undo the swap.
```

Reference 1. K-means and K-medoids: applet

Assumptions made by K Means

Simple and straightforward explanation with figures found at mbmlbook.

Summary:

1. All clusters are the same size.
2. Clusters have the same extent in every direction.
3. Clusters have similar numbers of points assigned to them.

Find a demonstration at Demonstration of k-means assumptions — scikit-learn 1.0.1 documentation

Implementation of K Means

sklearn.cluster.KMeans — scikit-learn 1.0.1 documentation

Sample Implementation on the IRIS Dataset

From Scratch Implementation - K Means Clustering | K Means Clustering Algorithm in Python