Indian Institute of Technology, Kanpur
Machine Learning Lecture Scribes, 17–18

Instructors: Piyush Rai and Purushottam Kar
Authors: Gurpreet Singh
Date: September 8, 2017

LECTURE
1

Clustering and K-Means Algorithm

1 Clustering

Definition 1 *Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.*
[1]

A cluster essentially represents a set of samples which are visually or statistically similar. Similarity can be defined using various methods, one such metric is explained in the k-means algorithm.

Clustering is the process of separating samples into their various clusters. This is a very crucial method, and has various applications in the world of Machine Learning and Data Mining.

2 K-Means Clustering

K-Means clustering partitions the sample data into K voronoi cells. The K-Means objective is based on Gaussian based kernels, with the similarity/closeness measure as the euclidian distance *i.e.* the l_2 norm of the difference of the point vectors.

Essentially, this is a NP hard problem, however there are heuristic algorithms, which are used to quickly converge the problem to a local optima. A very common algorithm is Lloyd's Algorithm (discussed in the next section), commonly known as the standard 'k-means algorithm'.

3 K-Means Algorithm

The K-Means clustering is a non-convex optimization problem, and is solved by alternating optimization *i.e.* we use an initial estimate of the cluster means to find cluster sets, which are further used to find a better solution. This process is repeated until convergence, and is essentially the K-Means algorithm.

The K-Means objective function tries to achieve minimum variance in the clusters. Formally, we need to assign clusters so that the sum of the variance multiplied by the cluster size in each cluster is minimum. This can be expressed as the following optimization problem

$$\arg\min_{S} \sum_{k=1}^{K} \sum_{\mathbf{x} \in S_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2 \tag{1}$$

It appears that a set S_k is represented by a single point $\boldsymbol{\mu}_k$. This is essentially the center point of the cluster k. Therefore, our objective is to find out $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$, as well as assign clusters to each point, such that the objective is minimized.

C_n represents the cluster of sample n and $\mathbf{x}_n$ represents the sample vector

The intuition behind this algorithm is that for any set of means (sets), any sample will be in the cluster with the mean of the cluster closest to it. This obviously reduces variance. If we, on the other hand, fix the cluster

Algorithm : *Alternating Optimization — K-Means Algorithm*

Initialize $\boldsymbol{\mu}_k$ to $\boldsymbol{\mu}_k^0$
while **no convergence**:

Update S

$$\forall n \in [N]: \qquad C_n = \arg\min_k \|\mathbf{x}_n - \boldsymbol{\mu}_k\|$$

Update $\boldsymbol{\mu}$

$$\forall n \in [K]: \qquad \boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\mathbf{x} \in S_k} \mathbf{x}$$

assignments, we can say that the point (mean) which will minimize the variance based on gaussian kernel will be the emperical mean of the data itself.

This is a hard assignment K-Means algorithm *i.e.* we only consider the possibility of a point being in one and only one cluster. Another approach to this is a soft assignment, in which we consider the expected probability of a point being in a cluster, and use alternating optimization then.

4 Probabilistic look on Lloyd's Algorithm

Consider the clusters to be represented by Gaussian Kernels with means $\boldsymbol{\mu}_k$ and variance σ^2 (Note, the variance is the same for all clusters, which is an essential assumption for K-Means Clustering). We then maximize the total probability of the points belonging to their corresponding clusters. Let us write the probability of a point being in the k^{th} cluster

$$P(C_n = k \,|\, \boldsymbol{\mu}_k) = \mathcal{N}\big(x_n \,|\, \boldsymbol{\mu}_k, \sigma^2\big)$$
$$\implies P(S \,|\, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \prod_{n=1}^N \mathcal{N}\big(x_n \,|\, \boldsymbol{\mu}_{C_k}, \sigma^2\big)$$

Since we need to maximize this probability, we maximize it's log

$$\max \quad \log\big(\mathcal{N}\big(\mathbf{x}_n \,|\, \boldsymbol{\mu}_{C_n}, \sigma^2\big)\big)$$
$$\implies \min \quad \frac{1}{\sigma^2} \sum_{n=1}^N \big\|\mathbf{x}_n - \boldsymbol{\mu}_{C_n}\big\|_2^2$$

This is the same optimization problem as given earlier. The probabilistic approach gives more insight into the working of the K-Means Clustering.

5 Problems in K-Means Clustering

K-Means Clustering works only for clusters which represent gaussian distributions. Hence, we cannot use K-Means Clustering for finding complex clusters or non-convex clusters.

The K-Means Algorithm is very sensitive to initialization, and hence one must be careful while initializing the cluster means.

The Algorithm can get stuck at a local optima, finding clusters different from those originally wanted. This is also a factor affected by the initialization of the cluster means.

References

[1] Cluster Analysis *Wikipedia* `https://en.wikipedia.org/wiki/Cluster_analysis`

[2] K-Means Clustering *Wikipedia* `https://en.wikipedia.org/wiki/K-means_clustering`

[3] C. Bishop. *Pattern Recognition and Machine Learning*

[4] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning — From Theory to Algorithms*