读书报告

09118119 黄一凡

# 自己提出的问题

#### 1.是否可以通过多次重复k-means算法，每次选取不同的数据点作为初始centroids，最后取最小SSE，来对算法进行优化？

可以。对于每一次选取不同的初始点，得到的只会是极小值，也就是局部最优值，如此重复几次后，虽然还是无法得到全局最优值， 但是相对只做一次可以对结果有一定的优化。

#### 2.对于出现outlier的情况，书上介绍的两种方法感觉是直接忽略它们，这样会影响最终cluster的形成吗，是否可以增加k的数量，也就是把outlier也作为一个聚类？

不会，因为在k-means算法中，不会只采用一个k进行试验，而是会采用许多k，最后根据其聚类的表现，选取最为合适的k值。 因此，可能在目前的k值下，该outlier会被忽略，但是在别的k值下，这个outlier又会被归到聚类中。

# 别人提出的问题

#### 1.使用k-means算法的时候，为什么全局最小值对于大规模数据集来说在计算上是不可行的？

因为每一次选取的初始点不同，所以最后生成的聚类结果也不同，因此每一次生成的结果只是基于目前初始点的最优值，也就是局部 最优值，而不是全局最优值。 对于大规模数据来说，因为其数据量非常大，初始点的选取不可能遍历所有的点，因此只会得到若干个局部最优值，而得不到 全局最优值。

#### 2.为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

因为这个点是潜在的可能性最大的聚类中心，可以有效减少后面的迭代次数，提高算法效率。

#### 3.对于4.2.3节中提出的计算good initial seed的方法，是否存在除了outliers以外的其它弊端，比如之后选取的seed和之前选取的seed重合或者距离很近？

这种特殊情况的确存在，但是这种情况可以通过后面的迭代过程来解决，只是在一定程度上影响了算法的效率。

#### 4.如果一个聚类任务的数据中，其属性既有离散值又有连续值怎么办？是否可以将连续属性按照数据集中的数据离散化分为几个区间，将区间视为离散的继续做？

如果在一个聚类任务中，既有离散值又有连续值并不会影响计算，因为它们都是可以计算的numerical value，只需要将它们按照距离公式 计算即可。

#### 5.To be safe, we may want to monitor these possible outliers over a few iterations and then decide whether to remove them. It is possible that a very small cluster of data points may be outliers. Usually, a threshold value is used to make the decision.其中，threshold value指的是什么？能具体讲一下怎么监视与这个方法的整体吗？

经过讨论，我们认为这里的阈值是针对某一点到其他所有centroid的距离。如果这个距离大于阈值，则认为该点为outlier。 监视过程即为在几个连续迭代周期中，若某一点一直达到阈值要求，则将其判断为outlier。

# 读书计划

#### 本周所读：

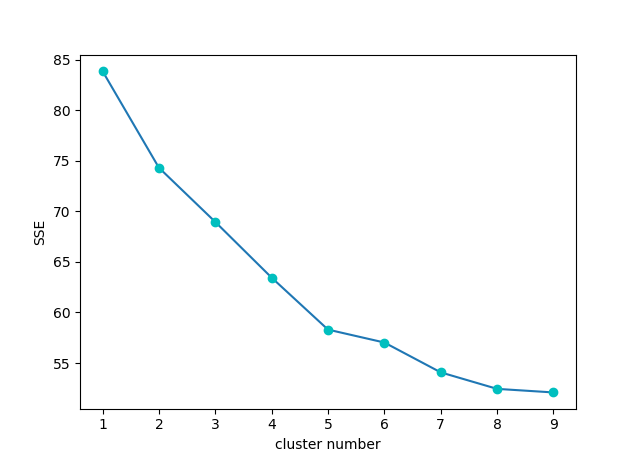
4.1-4.3

#### 下周计划：

第五章

# 代码实现

我根据书中伪代码具体实现了k-means算法，并用手肘法确定k的取值，下面是程序运行结果的截图：



# 读书摘要

下面是我读书时做的一些笔记整理：

# 4 Unsupervised Learning

In some applications, the data has no class attributes. The user wants to explore the data to find some intrinsic structures in them.

Clustering is one technology for finding such structures. It organizes data instances into similarity groups, called clusters such that the data instances in the same cluster are similar to each other and data instances in different clusters are very different from each other.

## 4.1 Basic Concepts

In the clustering literature, a data instance is also called a data point as it can be seen as a point in an -dimension space, where is the number of attributes in the data.

There are two types of clustering: partitional and hierarchical.

clustering needs a similarity function to measure how similar two data points are, or alternatively a distance function to measure the distance between two data points.

## 4.2 K-means Clustering

The k-means algorithm is the best known partitional clustering algorithm.

### 4.2.1 K-means Algorithm

Let the set of data point be:

The k-means algorithm partitions the given data into k clusters. Each cluster has a cluster center, which is also called the cluster centroid.

The centroid, usually used to represent the cluster, is simply the mean of all the data points in the cluster, which gives the name to the algorithm.

The stopping (or convergence) criterion can be any one of the following:

1. no (or minimum) re-assignments of data points to different clusters.
2. no (or minimum) change of centroids.
3. minimum decrease in the sum of squared error (SSE)

One problem with the k-means algorithm is that some clusters may become empty. To deal with an empty cluster, we can choose a data point as the replacement centroid.

### 4.2.2 Disk Version of the K-means Algorithm

The k-means algorithm may be implemented in such a way that it does not need to load the entire data set into the main memory, which is useful for large data sets.

The whole clustering process thus scans the data t times, where t is the number of iterations before convergence, which is usually not very large (< 50).

Although there are several special algorithms that scale-up clustering algorithms to large data sets, they all require sophisticated techniques.

### 4.2.3 Strengths and Weaknesses

The main strengths of the k-means algorithm are its simplicity and efficiency.

Its time complexity is , where is the number of data points, is the number of clusters, and is the number of iterations.

The weaknesses and ways to address them are as follows:

1. The algorithm is only applicable to data sets where the notion of the mean is defined. Thus, it is difficult to apply to categorical data sets. There is, however, a variation of the k-means algorithm called k-modes, which clusters categorical data.
2. The user needs to specify the number of clusters k in advance. In practice, several k values are tried and the one that gives the most desirable result is selected. We will discuss the evaluation of clusters later.
3. The algorithm is sensitive to outliers. Outliers are data points that are very far away from other data points. Outliers could be errors in the data recording or some special data points with very different values.
   * There are several methods for dealing with outliers
   * One simple method is to remove some data points in the clustering process that are much further away from the centroids than other data points
   * Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small. We can use the sample to do a pre-clustering and then assign the rest of the data points to these clusters
4. We can use the sample to do a pre-clustering and then assign the rest of the data points to these clusters.
5. The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids.

## 4.3 Representation of Clusters

Once a set of clusters is found, the next task is to find a way to represent the clusters.

### 4.3.1 Common Ways of Representing Clusters

There are there main ways to represent clusters: 1. Use the centroid of each cluster to represent the cluster. 2. Use classification models to represent clusters. 3. Use frequent values in each cluster to represent it. This method is mainly for clustering of categorical data. It is also the key method used in text clustering, where a small set of frequent words in each cluster is selected to represent the cluster.

### 4.3.2 Clusters of Arbitrary Shapes

Hyper-elliptical and hyper-spherical clusters are usually easy to represent.

However, other arbitrary shaped clusters are hard to represent especially in high dimensional spaces.

Due to the difficulty of representing an arbitrarily shaped cluster, an algorithm that finds such clusters may only output a list of data points in each cluster, which are not as easy to use. These kinds of clusters are more useful in spatial and image processing applications, but less useful in others.