K均值聚类理论与实践

- 什么是聚类分析
- K均值聚类分析的原理
- 如何进行K均值分析
- 如何评估聚类分析的效果

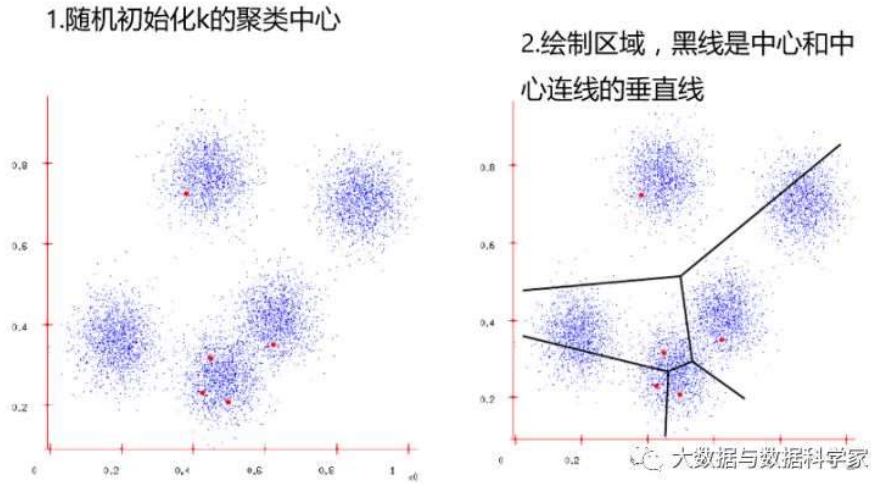
1 聚类分析

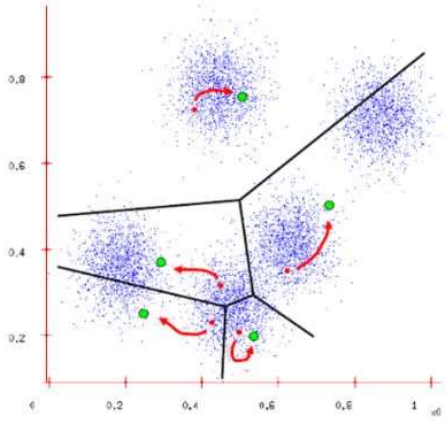
聚类分析是将样品或变量按照它们在性质上的亲疏程度进行分类的数据分析方法。聚类分析是典型的无监督分析方法，也就是没有关于样品或变量的分类标签，分类需要依据样品或者变量的亲疏程度进行。而亲疏程度可以用个体间的差异性或者相似度来表示。

2 K均值聚类

k-means算法是一种很常见的聚类算法，它的基本思想是：通过迭代寻找k个聚类的一种划分方案，使得用这k个聚类的均值来代表相应各类样本时所得的总体误差最小，即最小化代价函数：

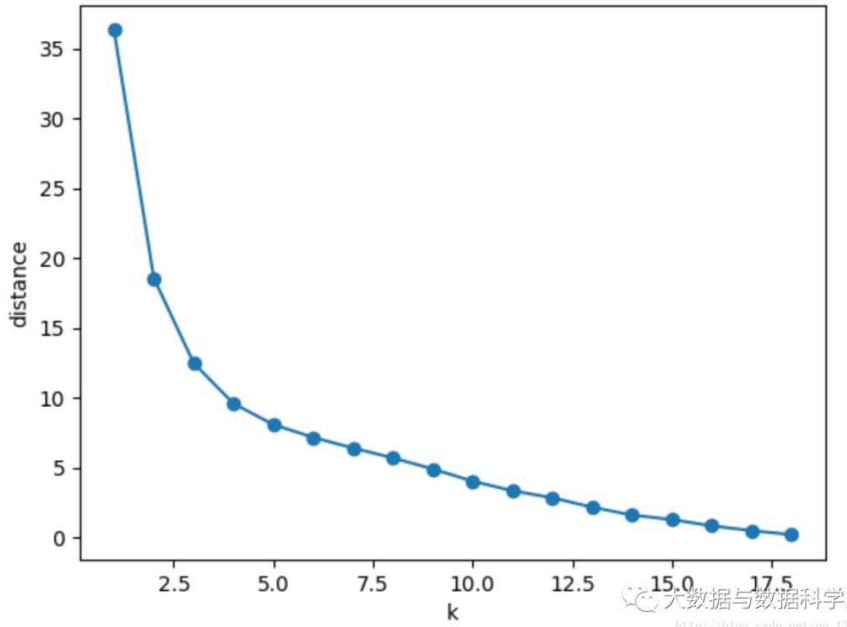
$$J(c,\mu)=\sum_{i=1}^{k}\| x^{(i)}-\mu_{c^{(i)}} \|^2$$





3.计算划分区域内所有点的均值，作为新的中心点
反复迭代这个过程，直到结果停止变换

那么该如何选择比较好的聚类数呢？答案是用肘部法则去确认。肘部法则通过把不同聚类数的代价函数算出来，作出曲线，比如下图，k取5的时候是比较合适的，取值过大会增加计算量，取值过小则模型质量不好。



3 K均值分析实例

原理了解得差不多了，我们来试着动手操作一下吧！以下数据是针对美国某区域的一次人口普查结果，数据来源为http://archive.ics.uci.edu/ml/datasets/Adult。

age	workclass	fnlwgt	education	education_num	maritial_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K

50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K

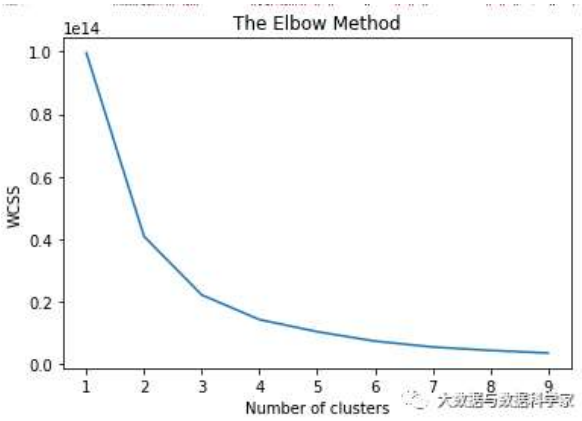
我们先处理数据，将分类数据转换为数值数据。

```
dataset = pd.read_excel('ch05_adult.data.xls')
dataset=dataset.dropna()
X=pd.get_dummies(dataset)
```


然后用肘部法则找出最佳的聚类数。

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 20):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X2)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 20), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```


结果如下:



可以看出，聚类数为3或者4时都比较好，我们选取4为聚类数。然后执行k-means算法就可以啦!

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 4, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)
```


4 评估聚类分析效果

我们经常用CH指标来评估聚类分析的效果。CH指标通过类内离差矩阵描述紧密度，类间离差矩阵描述分离度，指标定义如下：

$$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)}$$

其中，n表示聚类的数目 ,k表示当前的类,trB(k)表示类间离差矩阵的迹,trW(k)表示类内离差矩阵的迹。CH越大代表着类自身越紧密，类与类之间越分散，即聚类结果更优。

算法实现如下：

```
kmeans_model = kmeans.fit(X)
labels = kmeans_model.labels_
from sklearn import metrics
metrics.calinski_harabaz_score(X, labels)
```


当然，想要得到更好的模型，不一定把所有的指标都用上，比如本数据集中的education和education_num两个指标是强烈正相关的，不需要两个都用上。小伙伴们可以多试验几组指标，看看哪组可以得到更好的效果哦。

原创： 莫心瑶、江欣恺 大数据与数据科学家