一、简介

 K均值聚类算法是先随机选取K个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

二、

1.便于理解，首先创建一个明显分为2类20\*2的例子（每一列为一个变量共2个变量，每一行为一个样本共20个样本）：

import numpy as np

c1x=np.random.uniform(0.5,1.5,(1,10))

c1y=np.random.uniform(0.5,1.5,(1,10))

c2x=np.random.uniform(3.5,4.5,(1,10))

c2y=np.random.uniform(3.5,4.5,(1,10))

x=np.hstack((c1x,c2x))

y=np.hstack((c2y,c2y))

X=np.vstack((x,y)).T

print(X)

结果：

[[1.4889993 4.18741329]

[0.73017615 4.07842216]

[1.15522846 4.05744838]

[1.40768457 3.76674812]

[1.376212 3.95063903]

[1.20821055 4.34138767]

[0.73898392 3.55026013]

[0.97116627 3.65432314]

[0.98267302 4.16731561]

[1.06346541 4.44383585]

[4.10945954 4.18741329]

[3.75288064 4.07842216]

[4.29638229 4.05744838]

[3.95221785 3.76674812]

[4.09826192 3.95063903]

[4.04840874 4.34138767]

[4.29594009 3.55026013]

[3.56931245 3.65432314]

[3.57962941 4.16731561]

[3.65208848 4.44383585]]

1. 引用Python库将样本分为两类（k=2），并绘制散点图：

#只需将X修改即可进行其他聚类分析

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

kemans=KMeans(n\_clusters=2)

result=kemans.fit\_predict(X) #训练及预测

print(result) #分类结果

plt.rcParams['font.family'] = ['sans-serif']

plt.rcParams['font.sans-serif'] = ['SimHei'] #散点图标签可以显示中文

x=[i[0] for i in X]

y=[i[1] for i in X]

plt.scatter(x,y,c=result,marker='o')

plt.xlabel('x')

plt.ylabel('y')

plt.show()

结果：

[0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1]

1. 如果K值未知，可采用肘部法选择K值（假设最大分类数为9类，分别计算分类结果为1-9类的平均离差，离差的提升变化下降最抖时的值为最优聚类数K）：

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from scipy.spatial.distance import cdist

K=range(1,10)

meanDispersions=[]

for k in K:

kemans=KMeans(n\_clusters=k)

kemans.fit(X)

#计算平均离差

m\_Disp=sum(np.min(cdist(X,kemans.cluster\_centers\_,'euclidean'),axis=1))/X.shape[0]

meanDispersions.append(m\_Disp)

plt.rcParams['font.family'] = ['sans-serif']

plt.rcParams['font.sans-serif'] = ['SimHei'] #使折线图显示中文

plt.plot(K,meanDispersions,'bx-')

plt.xlabel('k')

plt.ylabel('平均离差')

plt.title('用肘部方法选择K值')

plt.show()

1. 实例分析（对某网站500家饭店价格及评论进行聚类）

import numpy as np

from sklearn.cluster import KMeans

from scipy.spatial.distance import cdist

import matplotlib.pyplot as plt

import pandas as pd

data=pd.read\_excel('data.xlsx',header=0).iloc[:501,3:5]

per\_25=data.describe().iloc[4,1]

per\_75=data.describe().iloc[6,1]

data=data[(data.iloc[:,1]>=per\_25)&(data.iloc[:,1]<=per\_75)] #选择位于四分位数之内的数

X=np.array(data)

K=range(1,10)

meanDispersions=[]

for k in K:

kemans=KMeans(n\_clusters=k)

kemans.fit(X)

meanDispersions.append(sum(np.min(cdist(X,kemans.cluster\_centers\_,'euclidean'),axis=1))/X.shape[0])

plt.rcParams['font.family'] = ['sans-serif']

plt.rcParams['font.sans-serif'] = ['SimHei']

plt.plot(K,meanDispersions,'bx-')

plt.xlabel('k')

plt.ylabel('平均离差')

plt.title('用肘部方法选择K值')

plt.show()

具体聚类过程

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

kemans=KMeans(n\_clusters=3)

result=kemans.fit\_predict(X)

print(result)

x=[i[0] for i in X]

y=[i[1] for i in X]

plt.scatter(x,y,c=result,marker='o')

plt.xlabel('avgPrice')

plt.ylabel('llCommentNum')

plt.title('对500家饭店价格与评论数进行聚类')

聚类结果：

[2 0 0 0 0 1 0 0 2 0 0 2 1 2 0 1 2 0 2 2 2 0 0 0 0 1 2 0 1 0 0 2 2 2 2 2 2

2 2 0 1 0 0 0 1 0 2 2 0 2 2 0 0 2 2 2 1 0 1 1 1 0 0 0 0 1 2 1 2 0 2 1 0 0

2 1 1 0 0 1 2 2 0 2 2 1 0 2 1 0 2 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 0 2 1 2 1

1 0 0 1 0 1 2 1 0 1 1 0 1 1 0 1 0 2 1 1 0 1 0 2 0 2 1 2 1 1 0 0 1 0 1 0 1

0 2 0 1 1 0 1 0 0 1 1 1 1 0 0 0 0 1 0 0 0 2 0 1 1 0 1 0 1 0 0 0 0 1 1 0 1

2 0 1 1 2 0 1 0 0 1 1 1 1 1 0 0 0 1 1 1 2 0 1 1 1 2 2 0 0 2 1 1 2 1 1 1 0

1 1 0 1 2 2 0 2 2 2 0 1 0 1 1 2 1 1 1 0 1 1 1 1 0 0 0 0 1]