Практическая работа № 4

Для заданного набора исходных данных провести кластеризацию

* методом иерархического кластерного анализа
* методом к-средних.

Часть 1. Порядок работы (иерархический кластерный анализ)

1. Импортировать требуемые библиотеки
2. Просмотреть заголовки набора данных
3. Просмотреть содержимое набора данных, если возможно.
4. Провести кластеризацию с помощью команды

link=linkage(df,’ward’,’euclidean’)

Просмотреть содержимое объекта link . Объект link – матрица, каждая строка которой результат слияния очередной пары кластеров с номерами

link[i,0] и link[i,1]. Новому кластеру присваивается номер n+1. Строка link[i,2] содержит расстояние между объединяемыми кластерами, link[i,3]

размер кластера.

1. Вывести дендрограмму

dn= dendrogram(link, orientation=’right’)

1. Найти среднее арифметическое всех показателей по каждому кластеру

df.groupby(‘cluster’).mean()

1. Определить сколько наблюдений попало в каждый кластер

df.groupby(‘cluster’). size()

Часть 2. Порядок работы (метод К-средних)

1. Импортировать требуемые библиотеки
2. Просмотреть заголовки набора данных
3. Просмотреть содержимое набора данных, если возможно.
4. Провести кластеризацию для количества кластеров к=2

From sklearn.cluster import KMeans

инициализация модели

model = KMeans(n\_clusters=2, random\_state=42)

подгонка модели по данным из df

model. fit(df)

KMeans(algorithm=’auto’, copy\_x=true, init=’K-means++’, max\_iter=300, random\_state=42, tol=0,0001,verbose=0)

просмотреть результаты кластеризации

array(…..)

координаты центров кластеров

model. cluster\_centers\_

предсказание для новых наблюдений-определить к какому кластеру относятся новые наблюдения

predict new\_items[…..]

model.predict(new\_items)