Кластеризация

Кластеризация

In
```
from sklearn.cluster import KMeans

# n_clusters – число кластеров
# init – начальные центроиды
model = KMeans(n_clusters=n_clusters, init=centers, random_state=12345)
model.fit(data)

# Получение центроидов кластеров
print(model.cluster_centers_)
# значение целевой функции
print(model.inertia_)
```

Построение графика pairplot с заливкой кластеров и отображением центроидов

In
```
import pandas as pd
from sklearn.cluster import KMeans
import seaborn as sns

centroids = pd.DataFrame(model.cluster_centers_, columns=data.columns)
# Добавление столбца с номером кластера
data['label'] = model.labels_.astype(str)
centroids['label'] = ['0 centroid', '1 centroid', '2 centroid']
# Сброс индекса понадобится дальше
data_all = pd.concat([data, centroids], ignore_index=True)

# Построение графика
sns.pairplot(data_all, hue='label', diag_kind='hist')
```

Построение графика Pairgrid с заливкой кластеров, начальными и конечными центроидами

In
```
import pandas as pd
from sklearn.cluster import KMeans
import seaborn as sns

centroids = pd.DataFrame(model.cluster_centers_, columns=data.columns)
# Добавление столбца с номером кластера
data['label'] = model.labels_.astype(str)
centroids['label'] = ['0 centroid', '1 centroid', '2 centroid']
# Сброс индекса понадобится дальше
data_all = pd.concat([data, centroids], ignore_index=True)

# Построение графика
pairgrid = sns.pairplot(data_all, hue='label', diag_kind='hist')
pairgrid.data = pd.DataFrame([[20, 80, 8], [50, 20, 5], [20, 30, 10]], \
                columns=data.drop(columns=['label']).columns)
pairgrid.map_offdiag(func=sns.scatterplot, s=200, marker='*', color='red')
```

Поиск оптимального числа кластеров методом локтя

In
```
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

distortion = []
K = range(1, 8) # число кластеров от 1 до 7
for k in K:
    model = KMeans(n_clusters=k, random_state=12345)
    model.fit(data)
    distortion.append(model.inertia_)

plt.figure(figsize=(12, 8))
plt.plot(K, distortion, 'bx-')
plt.xlabel('Число кластеров')
plt.ylabel('Значение целевой функции')
plt.show()
```

Словарь

Кластеризация (clustering)
объединение похожих объектов в группы, или кластеры.

Центроид (centroid)
центр кластера.