Mini-projeto K-Means

Prática / Experimentos

ANACONDA®

Prática / Experimentos

sklearn.cluster.KMeans

class `sklearn.cluster.` **`KMeans`** (*n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto'*) [source]

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Prática / Experimentos

sklearn_extra.cluster.KMedoids

class **sklearn_extra.cluster.KMedoids**(*n_clusters=8, metric='euclidean', method='alternate', init='heuristic', max_iter=300, random_state=None*) [source]

k-medoids clustering.

Read more in the User Guide.

Parameters: **n_clusters : int, optional, default: 8**

The number of clusters to form as well as the number of medoids to generate.

metric : string, or callable, optional, default: 'euclidean'

What distance metric to use. See :func:metrics.pairwise_distances metric can be 'precomputed', the user must then feed the fit method with a precomputed kernel matrix and not the design matrix X.

method : {'alternate', 'pam'}, default: 'alternate'

Which algorithm to use. 'alternate' is faster while 'pam' is more accurate.

init : {'random', 'heuristic', 'k-medoids++', 'build'}, or array-like of shape

https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html html

Prática / Experimentos

Conjunto de Dados Dados da Faculdade de Notícias dos EUA e do Relatório Mundial

777 Observações, 18Atributos (17 Descritores, 1 Atributo de Classe Binária)
Objetivo: Identificar se uma faculdade é pública ou privada
Link: https://www.kaggle.com/flyingwombat/us-news-and-world-reports-college-data

Diretrizes

- O problema deve ser tratado de forma **não-supervisionada**
- **Compare os algortimos K-Means e K-Medoids (quando os clusters tiverem definidos, calcule as distâncias Intra-clusters e Inter-clusters que podem dar uma ideia de boa formação dos clusters.**
- As métricas **elbow method, Calinski-Harabasz, Davies-Bouldin, Silhouette e BIC devem ser usadas para indicar o melhor valor de k**.
- Disponíveis em (https://github.com/smazzanti/are_you_still_using_elbow_method/blob/main/are-you-still-using-elbow-method.ipynb). (https://towardsdatascience.com/are-you-still-using-the-elbow-method-5d271b3063bd).
- Quando tiver os clusters definidos, pode fazer “perguntas” em cada cluster. Por exemplo, como se distribuem os valores para a variável X em cada cluster. Qual o valor médio da variável Z em cada cluster. Entre outras.

Prática / Experimentos

https://colab.research.google.com/drive/1HghRqEyy6WO949smwB0a29rMExHgO_kY

Link Elbow Method: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a