Clusterização - K-Means e Mean Shift

Anderson Sergio Oyama - RA: 91804
Pedro Henrique Torres Peres Garozi - R.A.:90552

Universidade Estadual de Maringá
Departamento de Informática - DIN
Prof. Dr. Wagner Igarashi

janeiro de 2020

Cluster

Clustering, ou agrupamento, é uma forma de organizar os dados através de construção de clusters, conjuntos. Para a elaboração de conjuntos, será necessário definir critério para que o algoritmo possa se basear na análise e definir em qual conjunto o elemento X pertence. Vale ressaltar que a característica do elemento X no conjunto, possui forte semelhança com os demais elementos do conjunto.

Mean Shift

O *Mean Shift* é um algoritmo não paramétrico e não supervisionado utilizado para estimar o gradiente de uma função de probabilidade. Para isso, *Mean Shift* realiza a busca pelo máximo local da função de probabilidade $f(\vec{x})$. Para isso, é necessário realizar o calculo do gradiente $\nabla f(\vec{x})$ da função.
Uma característica do algoritmo é que, o centro de massa é deslocado em direção na maior variação de concentração de pontos. Além disso, o algoritmo não exige que seja conhecido o número ou o formato dos clusters.

K-Means

O K-means agrupa dados tentando separar amostras em n grupos de igual variância, minimizando um critério conhecido como a inércia. Este algoritmo requer que o número de clusters seja especificado previamente.
A vantagem do K-means é que é rápido, sua complexidade em algoritmos heuristicos (geralment os algoritmos de Lloyd ou de Elkan) é de O(k n T) onde: n = número de pontos; k = número de clusters; T = número de iterações.

Base de dados

Para que possamos aplicar os algoritmos estaremos utilizando a uma base de dados. A base de dados escolhido foi a *Clustering basic benchmark*, disponível em http://cs.joensuu.fi/sipu/datasets/. Estaremos utilizando a **S-sets** e a **A-sets**.

Execução dos algoritmos

```
import pandas as pd
dataset = pd.read_csv(datasetLocal, sep=” ”, header=None)
```

Mean Shift

```
from sklearn.cluster import MeanShift, estimate_bandwidth
bandwidth = estimate_bandwidth(dataset, quantile=mediumDistance,
n_samples=sample)
```

Mean Shift

ms = MeanShift(bandwidth=bandwidth, $bin_seeding = True, cluster_all = True)ms.fit(dataset)labels = ms.labels_cluster_centers = ms.cluster_centers_labels_unique = np.unique(labels)n_clusters_=len(labels_unique)$

Mean Shift

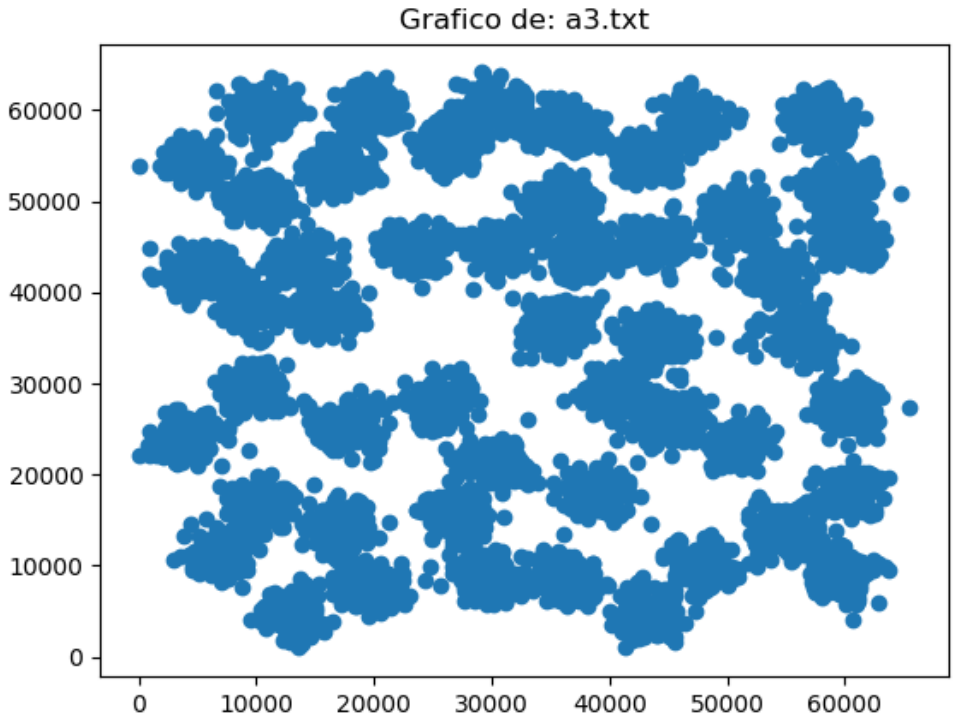
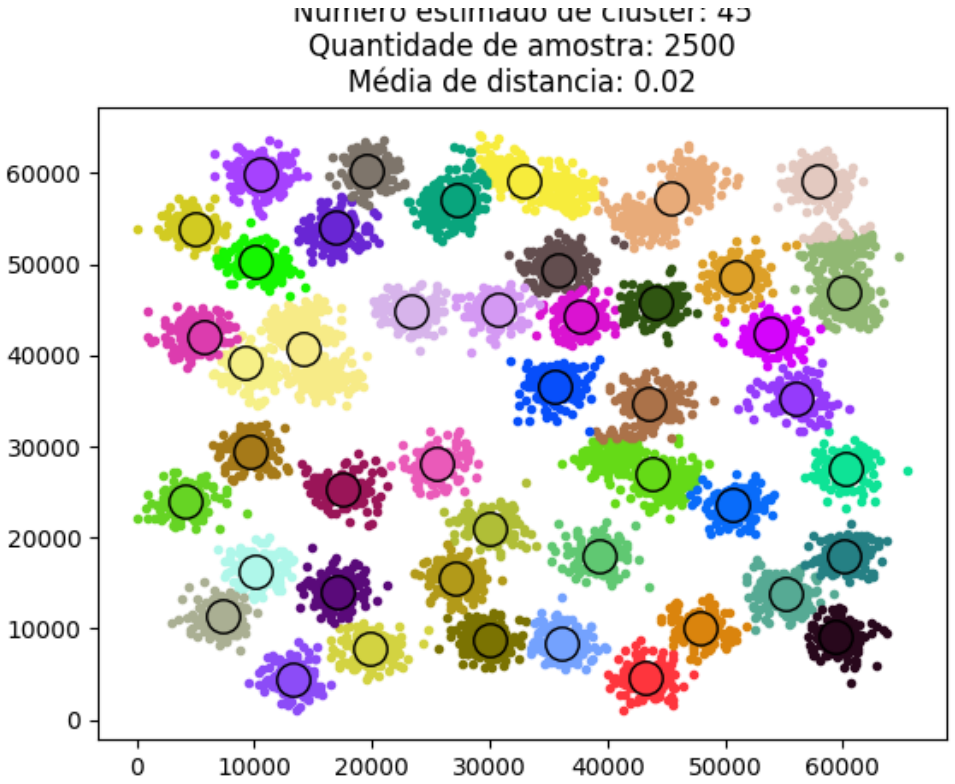


Grafico de: a3.txt
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000

Mean Shift


Número estimado de cluster: 45
Quantidade de amostra: 2500
Média de distancia: 0.02
60000
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000

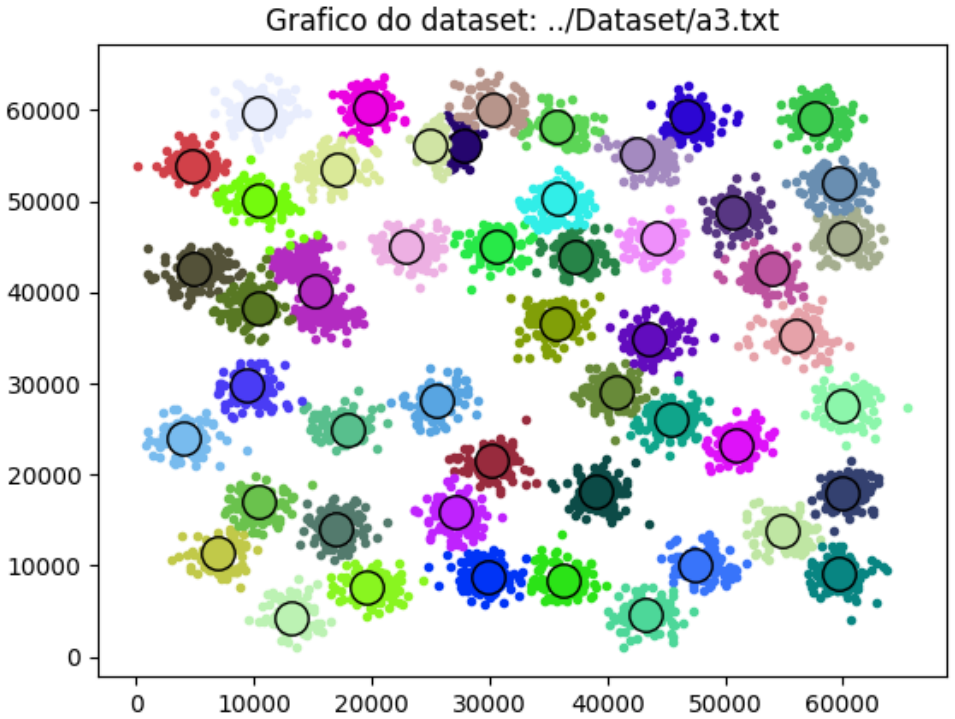
K-Means: Separação treino/testes

```
def split_sets(dataset):
    dataset = dataset.sample(frac=1)
    train_len = int(len(dataset) * 0.25)
    train_set = dataset.iloc[:train_len]
    test_set = dataset.iloc[train_len:]
    return train_set, test_set
```

K-Means: uso da biblioteca

```
kmeans = KMeans(n_clusters=int(sys.argv[2]), random_state=42).fit(train_set)
Y = kmeans.predict(test_set)
```

K-Means - Resultado



Referencia

Pasi Fränti and Sami Sieranoja. *K-means properties on six clustering benchmark datasets*. 2018. URL: http://cs.uef.fi/sipu/datasets/.

E. Rich. *The Gradual Expansion of Artificial Intelligence*. Vol. 17. University of Texas at Austin, 1984.