Introdução à Ciência de Dados

Prof. Francisco Rodrigues

Agrupamento de dados

1 - Considere os dados gerados pelo código abaixo. Usando o método k-means e a medida normalized mutual information, determine o número ideal de clusters para os dados abaixo. Veja o exemplo da aula.

```
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import normalized_mutual_info_score
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(201)

# create blobs
c = [(-2,1),(0,0),(4,6),(5,1),(6,12)]
n=300
data = make_blobs(n_samples=n, n_features=2, centers=c, cluster_std=1, random_state=50)
X = data[0]
labels = data[1]
plt.scatter(X[:,0], X[:,1], c=labels, cmap='viridis', s=50, alpha=0.9)
plt.show(True)
```

2 - Repita a análise feita no notebook da aula para dados gerados usando a função make_circles (https://scikit-learn.org/stable/datasets/index.html#sample-generators). Varie o número de observações e veja como se comporta o agrupamento. Avalie usando Coeficiente de Silhueta.