

UNIDADE 3 - Medidas de validação de clusters

Marta Noronha

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
CURSO: CIÊNCIA DE DADOS
APRENDIZADO DE MÁQUINA II



SUMÁRIO

- 3.1 Medidas de coesão e medidas de separação
- 3.2 Medidas internas (entropia, compacidade)
- 3.3 Medidas externas (rand index, Informação Mútua Normalizada, F-measure)
- 3.4 Definindo o número de clusters: métodos da silhueta e do cotovelo

Tipos de medidas de validação

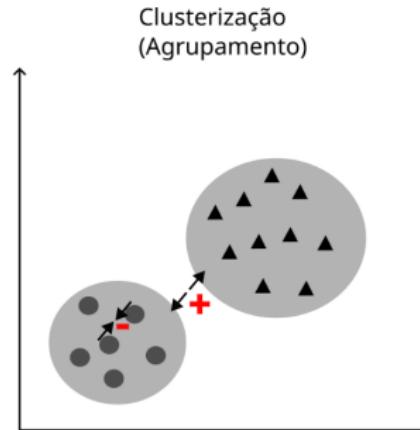
- **Medidas internas:** Usadas para medir a qualidade da estrutura de um cluster sem considerar informações externas.
- **Medidas externas:** Usadas para medir até que ponto o cluster e os rótulos correspondem aos rótulos de classe fornecidos externamente. Também chamado de validação por conhecimento externo.
- **Medidas relativas:** São usadas para comparar resultados obtidos por diferentes algoritmos de clusterização ou clusters diferentes.

Validação de clusters

- Clusterização é uma tarefa complexa onde se busca descobrir estruturas subjacentes no conjunto de dados.
- Descobrir a quantidade de grupos não é trivial.
- Em cenários onde os tipos de variáveis são mistas então as medidas não são efetivas [Liang et al. 2012].
- Problemas inerentes aos dados como dados ausentes e outliers podem tornar a análise mais complexa.
- A parametrização incorreta também pode favorecer a descoberta de clusters insignificantes ou sem uma interpretação para o domínio.

Medidas de coesão e separação

- Usadas para definir a qualidade dos clusters dado certo(s) critério(s).
- O critério é definido pela medida usada na avaliação intracluster e interclusters.
- Clusters devem possuir os objetos de seu cluster mais próximos (similares) entre si (coesão) e mais distantes dos objetos agrupados em clusters vizinhos (separação).



Medidas de coesão e separação

Coesão

- Coesão mede o quanto próximos estão os objetos dentro de um cluster.
- Pode ser calculada pela soma do erro quadrado médio dos objetos de cada cluster.
- Na fórmula, i é um dos clusters obtidos na clusterização, m_i é o centróide do cluster em análise e x é cada objeto em c_i .
- Quanto menor, melhor, por indicar que os clusters são mais coesos. (WSS -*Within Sum of Squares*)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

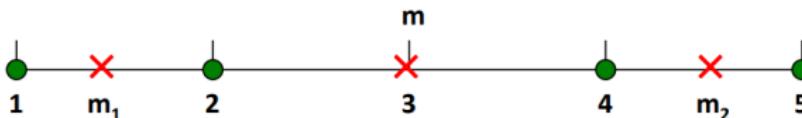
Medidas de coesão e separação

Separação

- Calculada pela soma do erro quadrado entre os centróides de cada cluster (BSS - *Between Sum of Squares*).
- Na fórmula, i é um dos clusters obtidos na clusterização, m_i é o centróide do cluster em análise e m é o centróide do conjunto de dados.
- Quanto maior, melhor, porque indica que os clusters estão bem separados.

$$BSS = \sum_i |C_i|(m - m_i)^2$$

Medidas de coesão e separação



$$K=1 : \quad WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

$$K=2 : \quad WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

$$K=4 : \quad WSS = (1-1)^2 + (2-2)^2 + (4-4)^2 + (5-5)^2 = 0$$

$$BSS = 1 \times (1-3)^2 + 1 \times (2-3)^2 + 1 \times (4-3)^2 + 1 \times (5-3)^2 = 10$$

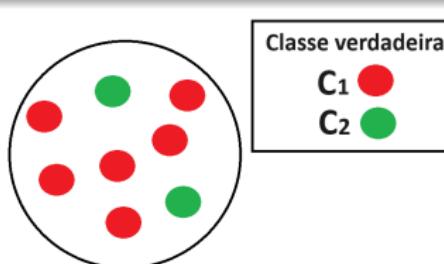
$$Total = 0 + 10 = 10$$

Fonte: https://cse.buffalo.edu/~jing/cse601/fa13/materials/clustering_basics.pdf

Medidas externas

Pureza de um único cluster

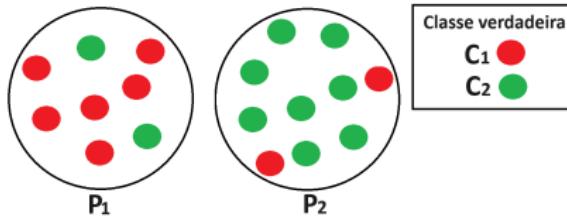
- A pureza do cluster é calculada quando se possui os índices dos objetos no cluster considerando o percentual da classe mais comum.
- Limitação: não considera a estrutura dos dados. Dois clusters podem ser considerados puros mesmo que um deles seja compacto e o outro seja disperso.
- Deve ser usada em conjunto com outras métricas de validação, como a entropia ou a silhueta, para obter uma avaliação mais completa da qualidade dos clusters.
- Exemplo: Um determinado grupo contém 8 objetos sendo que 6 são da classe majoritária. Logo a pureza do cluster é $6/8 = 0,75$ (equivale a 75%).



Medidas externas

Pureza de todos os clusters descobertos

- A pureza de todos os clusters é dada por $\frac{1}{N} \sum_k \max_j |C_k \cap P_j|$ sendo que C_k e P_j correspondem respectivamente a k-ésima verdadeira classe e ao j-ésimo cluster (ou partição P) descoberto - [Manning, Raghavan e Schütze 2008].
- Exemplo: Um cluster P_1 contém 8 objetos sendo 6 da classe majoritária. O cluster P_2 contém 10 objetos sendo 8 da classe majoritária. Logo a pureza do cluster é $\frac{1}{18} \times (6 + 8) = 0,778$ (equivale a 77.8%).



Medidas externas

Pureza: Prós

- Dados reais: Interpretação Intuitiva e sensibilidade ao desbalanceamento.
- Dados sintéticos: Avaliação Controlada sobre dados sintéticos com sensibilidade a estruturas controladas.

Pureza: Contras

- Dados reais: Depende de um cluster de referência e sensibilidade as diferentes formas e densidades de clusters pontuando igualmente clusters diferentes.
- Dados sintéticos: Simplificação excessiva em dados sintéticos pode não refletir a complexidade e a generalização sobre dados reais.

Medidas externas

Entropia

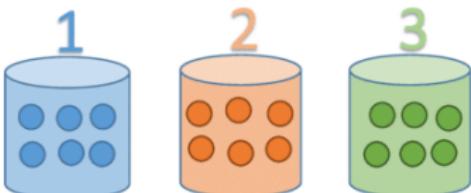
- Quantifica a impureza, mede a desordem dos clusters descobertos.
- Métrica usada para avaliar a qualidade do cluster em relação à distribuição das verdadeiras classes dos objetos.
- Quanto menor a entropia, mais homogêneo (puro) é o cluster.
- Frequentemente usada em algoritmos de árvore de decisão, mas também pode ser usada para validar os clusters quando os rótulos de classe existem.
- A entropia é de um cluster P é dada por $-\sum_{i=1}^k p_i \log_2 p_i$, onde k é o número de classes diferentes no cluster e p_i é a proporção de objetos que pertencem à classe i .
- Faixa: $\{0, 1\}$

Medidas externas

Entropia

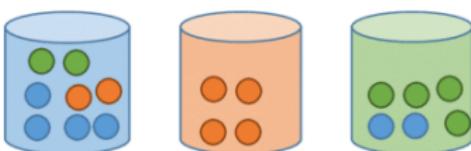
k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0

$$S_c = 0$$



k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	4/8	2/8	2/8	1.5
2	0	1	0	0
3	2/6	0	4/6	0.918

$$S_c = 0.971$$



k	p_{1k}	p_{2k}	p_{3k}	s_{Lk}
1	2/6	2/6	2/6	1.585
2	2/6	2/6	2/6	1.585
3	2/6	2/6	2/6	1.585

$$S_c = 1.585$$



$$S_c = \sum_1^K \frac{n_k}{N} s_{Lk} \quad : \quad s_{Lk} = \sum_1^L -p_{lk} \log_2 p_{lk}$$

Fonte: Mannheimer et al. 2019

Medidas externas

Entropia: Prós

- Dados reais: Interpretação Intuitiva e sensibilidade a estrutura de dados sendo ideal para medir a homogeneidade dos clusters.
- Dados sintéticos: Avaliação Controlada sobre dados sintéticos com sensibilidade a estruturas controladas.

Entropia: Contras

- Dados reais: Depende de um cluster de referência e sensibilidade à distribuição de classes em dados reais complexos.
- Dados sintéticos: Simplificação excessiva em dados sintéticos pode não refletir a complexidade e a generalização sobre dados reais.

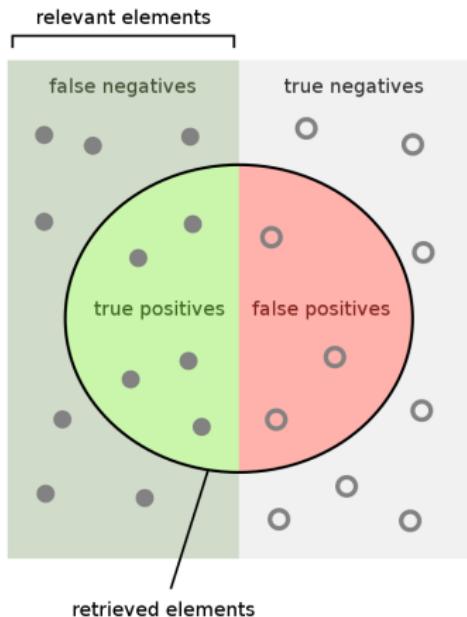
Medidas externas

F-Score

- Média harmônica entre a precisão e a sensibilidade.
- Precisão está relacionada a proporção de objetos corretamente agrupados no cluster em relação ao total de objetos do cluster.
- Sensibilidade é a proporção de objetos corretamente agrupados no cluster em relação ao total de objetos que realmente pertencem ao cluster.
- Faixa: {0, 1}
- $$F - Score = \frac{2 \times precisao \times sensibilidade}{precisao + sensibilidade}$$

Medidas externas

F-Score



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

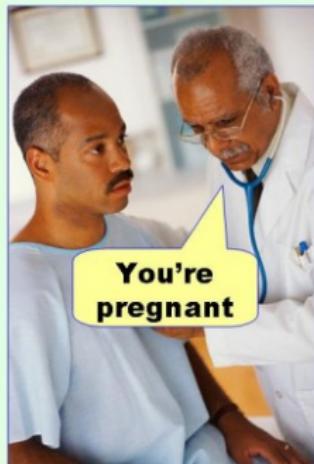
How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Medidas externas

F-Score

Type I error
(false positive)



Type II error
(false negative)



Medidas externas

F-Score: Prós

- Dados reais: Interpretação Intuitiva, Adequado para Dados Desbalanceados e Sensibilidade a Erros Tipo I e Tipo II
- Dados sintéticos: Avaliação Controlada sobre dados sintéticos e Adequado para Testes de comparação de algoritmos.

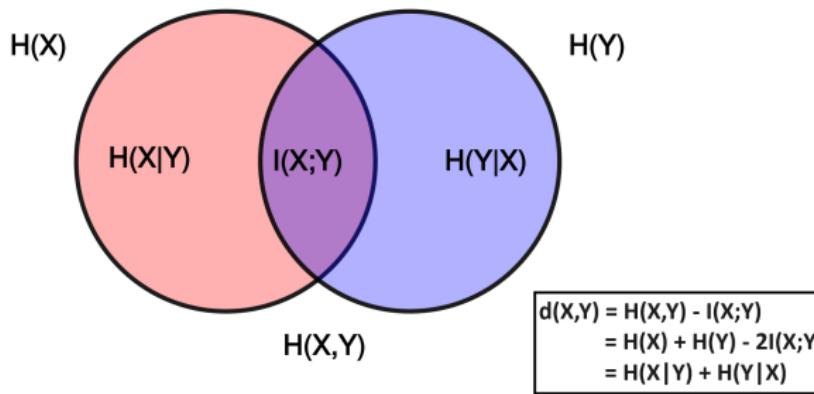
F-Score: Contras

- Dados reais: Dependência de Rótulos Verdadeiros e Sensibilidade a Inicialização e Hiperparâmetros dos algoritmos.
- Dados sintéticos: Simplificação excessiva em dados sintéticos pode não refletir a complexidade e a generalização sobre dados reais.

Medidas externas

Variação da informação

- Quantifica a diferença entre dois clusters (como o cluster sintético e o cluster descoberto) do conjunto de dados.
- Mede a quantidade de informação compartilhada entre dois clusters e a quantidade de informação exclusiva de cada cluster.
- Limitada inferiormente em 0.



Medidas externas

Variação da informação: Prós

- Dados reais: Sensibilidade a pequenas diferenças entre clusters.
- Dados sintéticos: Avaliação Controlada sobre dados sintéticos e Adequado para Testes de comparação de algoritmos.

Variação da informação: Contras

- Dados reais: Interpretação limitada por requerer conceitos de teoria de informação e dependência de um cluster de referência para cálculo.
- Dados sintéticos: Simplificação excessiva em dados sintéticos pode não refletir a complexidade e a generalização sobre dados reais.

Medidas externas

Medidas de validação defeituosas

- Medidas que produzem resultados de validação enganosos para K-means em dados com distribuições de classes distorcidas (assimetria das caudas).
- K-means tende a produzir clusters com tamanhos relativamente uniformes.
- Usa-se o coeficiente de variação, uma medida adimensional, para medir o grau de dispersão de uma distribuição aleatória, onde $CV = s/\bar{x}$, s é o desvio padrão do cluster e \bar{x} é a média do cluster.
- Quanto maior o CV mais disperso são os dados.

Medidas externas

Medidas de validação defeituosas

TABLE 23.3: Two Clustering Results

I	C_1	C_2	C_3	C_4	C_5
P_1	10	0	0	0	0
P_2	10	0	0	0	0
P_3	10	0	0	0	0
P_4	0	0	0	10	0
P_5	0	2	6	0	2

II	C_1	C_2	C_3	C_4	C_5
P_1	27	0	0	2	0
P_2	0	2	0	0	0
P_3	0	0	6	0	0
P_4	3	0	0	8	0
P_5	0	0	0	0	2

Fonte: Aggarwal e Reddy 2013

- Considere que a amostra contém 50 documentos de 5 classes. A dimensão das classes são 30, 2, 6, 10 e 2. Então o $CV_0 = 1,166$.
- CV_1 na tabela I é igual a 0, portanto os clusters são balanceados.
- CV_2 na tabela II é igual a 1,125.
- Considerando-se somente os CV 's pode-se dizer que o primeiro resultado é o correto, porém a distribuição da tabela II pelo CV é mais próximo do real.

EXERCÍCIOS

A matriz de confusão é dada por:

		Previsto	
		Positivo	Negativo
Real	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

1 - Sabendo que a precisão é dada por $P = \frac{VP}{VP+FP}$, a sensibilidade é dada por $S = \frac{VP}{VP+FN}$ e o F-Score é a média harmônica entre precisão e sensibilidade, ou seja, $F = \frac{2*P*S}{P+S}$, mostre a equação para o F-Score, em termos de VP, FP e FN .

EXERCÍCIOS

2 - Considere a seguinte tabela de contingência (verdadeiras classes ↓ × clusters descobertos →):

Classe \ Cluster	P_A	P_B	P_C
A	4	1	0
B	2	3	1
C	0	2	4
	S_{LA}	S_{LB}	S_{LC}

Total de objetos: $N = 17$.

- 1 Para cada cluster P_A , P_B e P_C , calcule:
 - 1 Pureza;
 - 2 Entropia (use $S_{LK} = -\sum_{i=1}^k p_i \log_2(p_i)$)
 - 3 F-Score (defina precisão e sensibilidade por classe e reporte o F-Score).
- 2 Calcule as métricas de pureza, entropia (use $E = \sum_1^k \frac{n_k}{N} * S_{LK}$) e F-Score para o conjunto total de clusters.

A pureza de todos os clusters é dada por $\frac{1}{N} \sum_k \max_j |C_k \cap P_j|$ sendo que C_k e P_j correspondem respectivamente a k-ésima verdadeira classe e ao j-ésimo cluster (ou partição P) descoberto.

EXERCÍCIOS

3 - Considere os seguintes conjuntos de pontos, representando dois clusters, em \mathbb{R}^1 :

$$C_1 = \{0, 2\}, \quad C_2 = \{5, 10\}.$$

Calcule:

- ① O centro dos dados $m = \frac{1}{N} \sum_i \sum_{x \in C_i} x$, onde N é o total de pontos do conjunto de dados;
- ② a coesão dada por $WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$;
- ③ e, a separação dada por $BSS = \sum_i |C_i| (m - m_i)^2$.
- ④ Interprete se os clusters são mais coesos ou mais separados.

EXERCÍCIOS

4 - Considere os seguintes conjuntos de pontos, representando três clusters, em \mathbb{R}^1 :

$$C_1 = \{1, 3\}, \quad C_2 = \{6, 7\}, \quad C_3 = \{9, 12\}.$$

Calcule:

- ① O centro dos dados $m = \frac{1}{N} \sum_i \sum_{x \in C_i} x$, onde N é o total de pontos do conjunto de dados;
- ② a coesão dada por $WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$;
- ③ e, a separação dada por $BSS = \sum_i |C_i| (m - m_i)^2$.
- ④ Interprete se os clusters são mais coesos ou mais separados.

EXERCÍCIOS

5 - Mostre o cálculo completo do coeficiente de variação (CV), apresentado em Aggarwal e Reddy 2013, para as classes verdadeiras (C_1 a C_5) e para cada Tabela.

TABLE 23.3: Two Clustering Results

I	C_1	C_2	C_3	C_4	C_5
P_1	10	0	0	0	0
P_2	10	0	0	0	0
P_3	10	0	0	0	0
P_4	0	0	0	10	0
P_5	0	2	6	0	2

II	C_1	C_2	C_3	C_4	C_5
P_1	27	0	0	2	0
P_2	0	2	0	0	0
P_3	0	0	6	0	0
P_4	3	0	0	8	0
P_5	0	0	0	0	2

- $CV = \frac{s}{\bar{x}} * 100$, onde s é o desvio padrão e \bar{x} é a média.
- Considere que a amostra contém 50 documentos de 5 classes. A dimensão das classes são 30, 2, 6, 10 e 2. Então o $CV_0 = 1,166$.
- CV_1 na tabela I é igual a 0.
- CV_2 na tabela II é igual a 1,125.

Medidas internas

- Serão abordadas medidas internas para clusters em que os objetos pertencem somente a um único cluster (*crisp cluster*).
- Medidas internas não utilizam informações externas como índices reais ou sintéticos para validar os resultados.
- Medidas internas se baseiam somente na estrutura dos clusters e em conceitos como a separação entre clusters e coesão interna de um cluster.

Medidas internas

Índice de silhueta (*Silhouette index*)

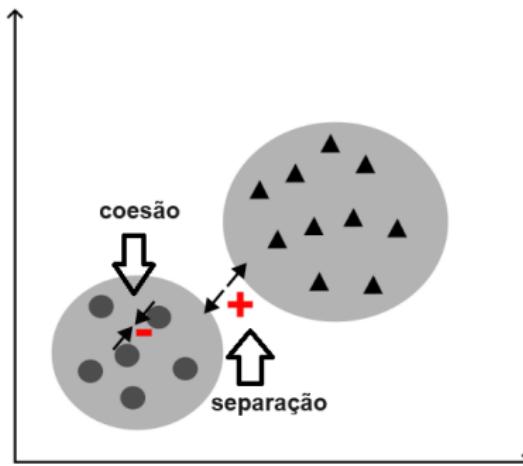
- Mensura o quanto próximo cada ponto em um cluster está dos pontos nos clusters vizinhos.
- Fornece análise visual dos parâmetros como o melhor número de clusters de tal forma que os clusters exibam maior coesão interna e maior separação entre os clusters.
- Faixa: {-1,1}.
 - 1: Os clusters estão bem separados uns dos outros e claramente distintos.
 - 0: Os clusters são indiferentes, ou a distância entre os clusters não é significativa.
 - -1: Os clusters foram atribuídos de maneira errada.
- A silhueta é bastante útil para validar os agrupamentos em conjuntos de dados com dimensões maiores do que 3 devido a falta de recursos para visualização.
- Útil para verificar o número ideal de clusters.

Medidas internas

Índice de silhueta

$$\text{Silhueta} = \frac{(b-a)}{\max(a,b)}$$

onde b é a separação entre os clusters e a é a coesão intracluster.



Medidas internas

Silhueta em python

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
%matplotlib inline
```

```
X= np.random.rand(70,2)
Y= 2 + np.random.rand(50,2)
W= 1 + np.random.rand(50,2)
Z= np.concatenate((X,Y,W))
Z=pd.DataFrame(Z) #converting into data frame for ease
```

Validação de clusters
oooooooo

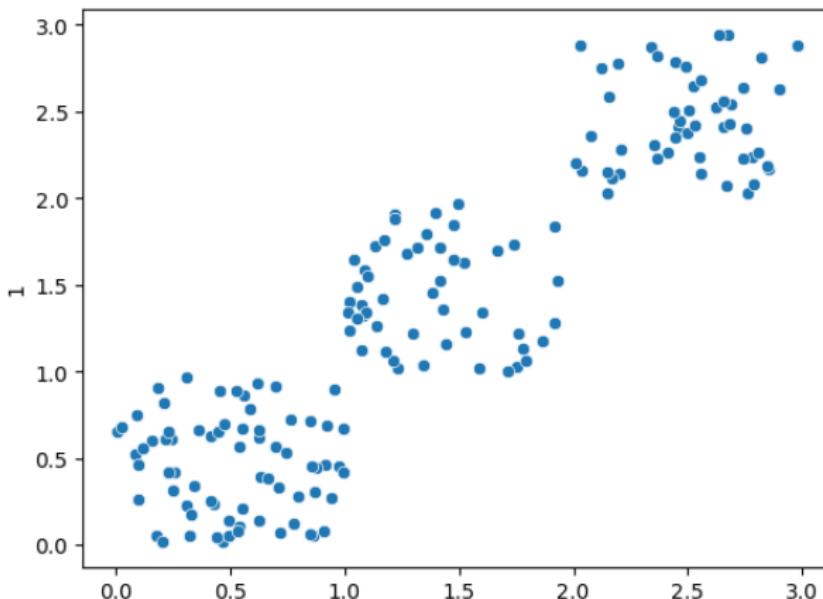
Medidas externas
oooooooooooooooooooo

Medidas internas
oooo●oooooooooooooooooooo

Medidas internas

Silhueta em python

```
sns.scatterplot(x=Z[0], y=Z[1])
```



Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooo●oooooooooooooooooooo

Medidas internas

Silhueta em python

```
KMean= KMeans(n_clusters=2)  
KMean.fit(Z)  
label=KMean.predict(Z)
```

```
print(f'Silhouette Score(n=2): {silhouette_score(Z, label)}')
```

```
Silhouette Score(n=2): 0.5072565233098857
```

```
sns.scatterplot(x=Z[0],y=Z[1],hue=label)
```

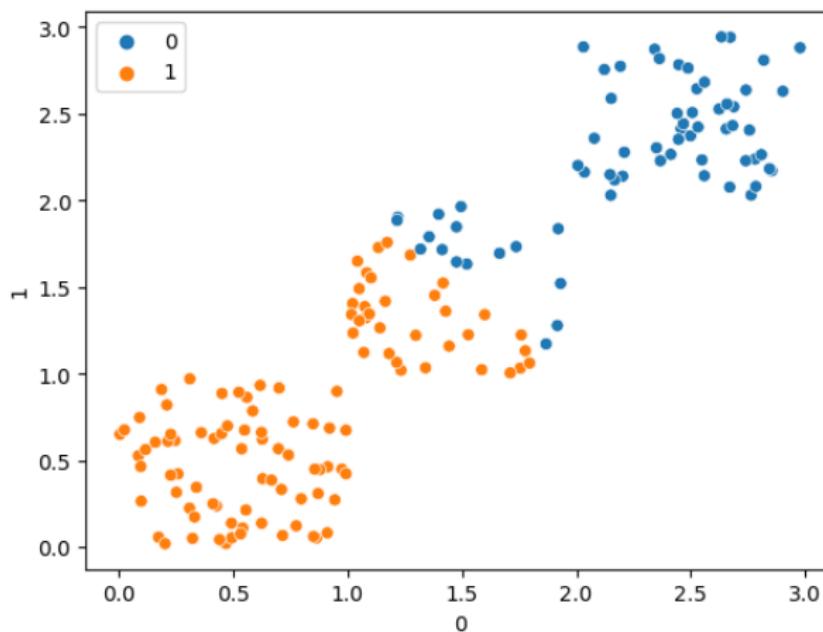
Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooooooooooooooooooo

Medidas internas

Silhueta em python (`n_clusters = 2`)



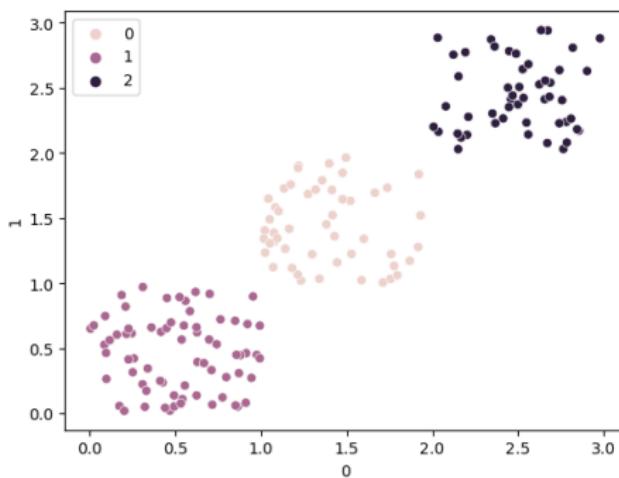
Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooo●oooooooooooooooooooo

Medidas internas

Silhueta em python (n_clusters = 3)



Silhouette Score(n=3): 0.6162241860864481

Validação de clusters
○○○○○○○

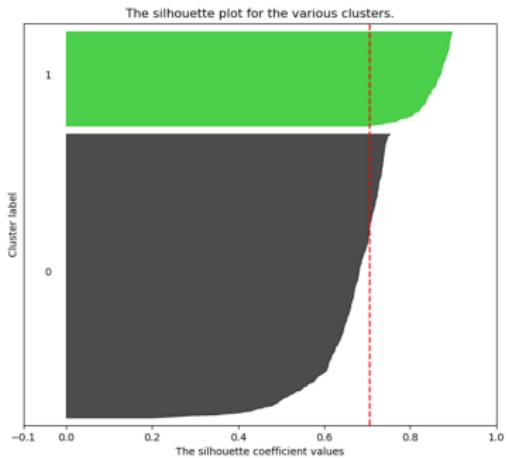
Medidas externas
○○○○○○○○○○○○○○○○○○○○○○

Medidas internas
○○○○○○●○○○○○○○○○○○○○○○○○○○○○○

Medidas internas

Visualização da silhueta sobre dados gerados por make-blobs

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



For n_clusters = 2 The average silhouette_score is : 0.7049787496083262

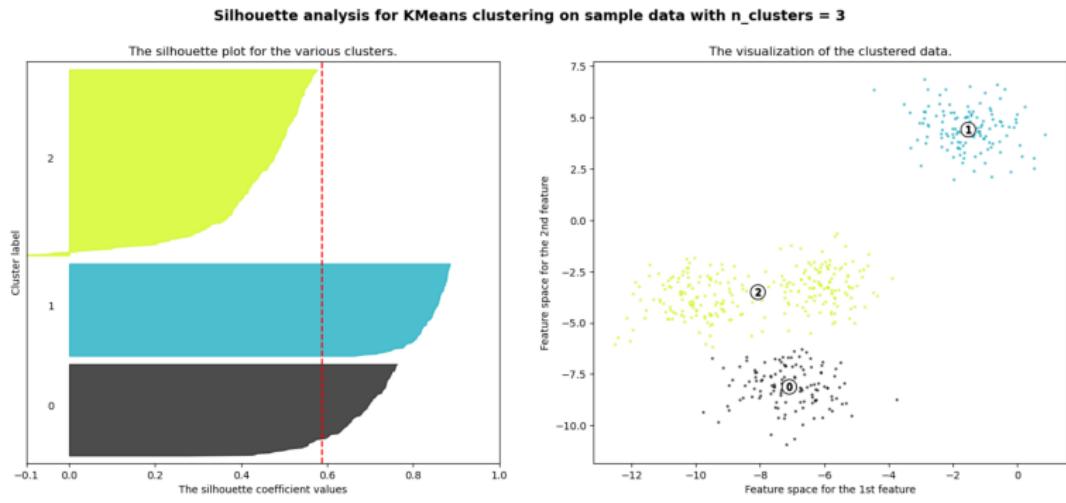
Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooooooo●oooooooooooooooooooo

Medidas internas

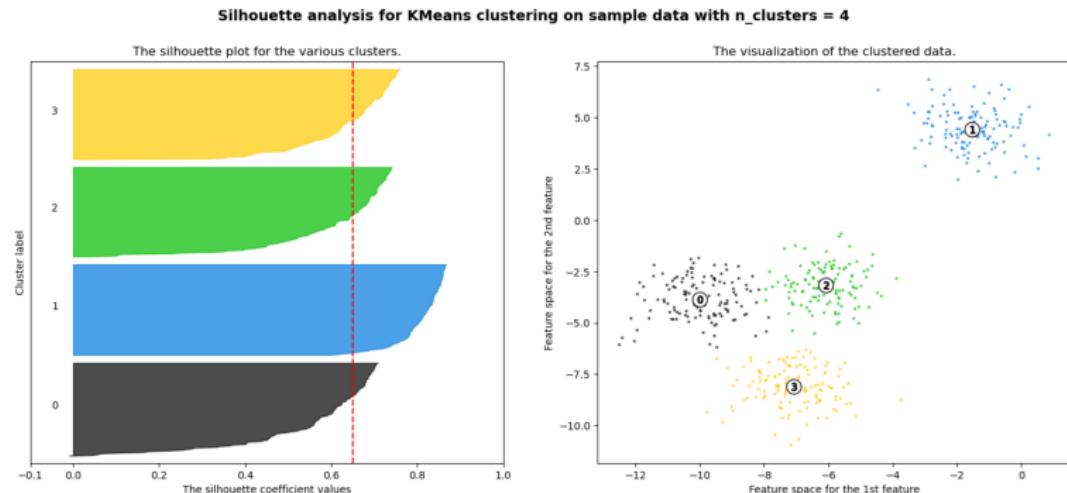
Visualização da silhueta sobre dados gerados por make-blobs



For n_clusters = 3 The average silhouette_score is : 0.5882004012129721

Medidas internas

Visualização da silhueta sobre dados gerados por make-blobs



For n_clusters = 4 The average silhouette_score is : 0.6505186632729437

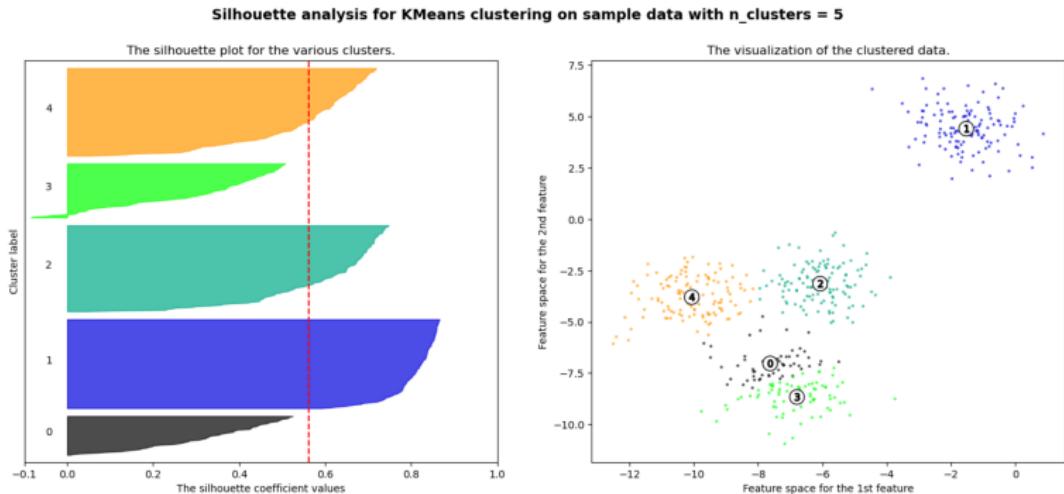
Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooooooo●oooooooooooo

Medidas internas

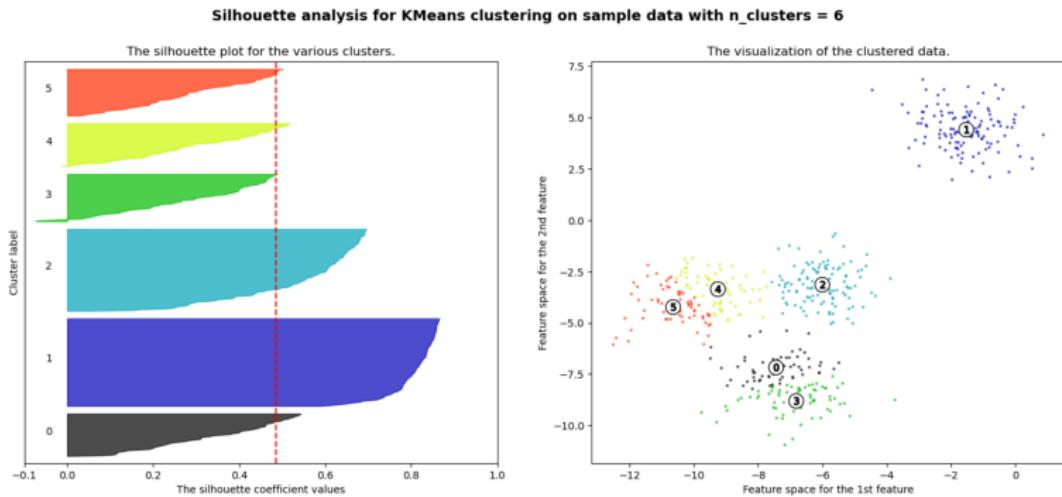
Visualização da silhueta sobre dados gerados por make-blobs



For n_clusters = 5 The average silhouette_score is : 0.561464362648773

Medidas internas

Visualização da silhueta sobre dados gerados por make-blobs



For n_clusters = 6 The average silhouette_score is : 0.4857596147013469

Validação de clusters
○○○○○○○

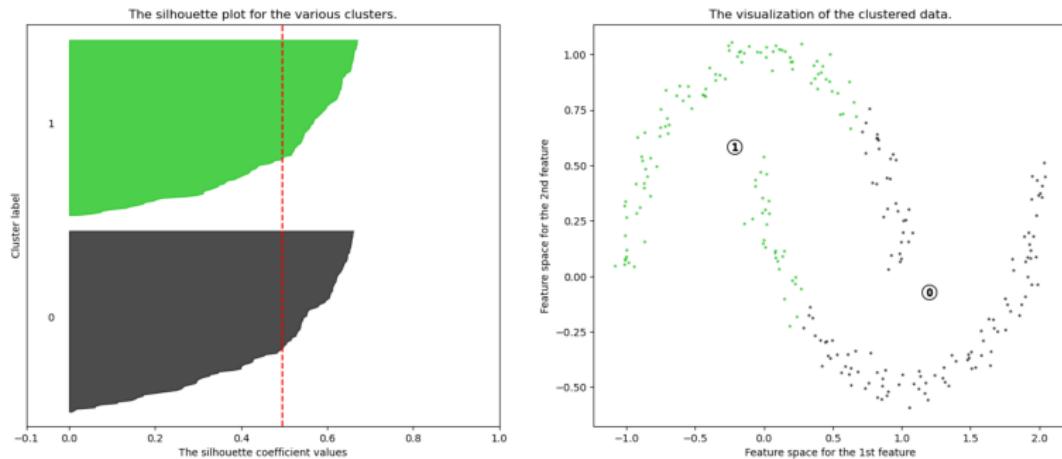
Medidas externas
○○○○○○○○○○○○○○○○○○

Medidas internas
○○○○○○○○○○●○○○○○○○○○○○○○○

Medidas internas

Visualização da silhueta sobre dados gerados por make-moons

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

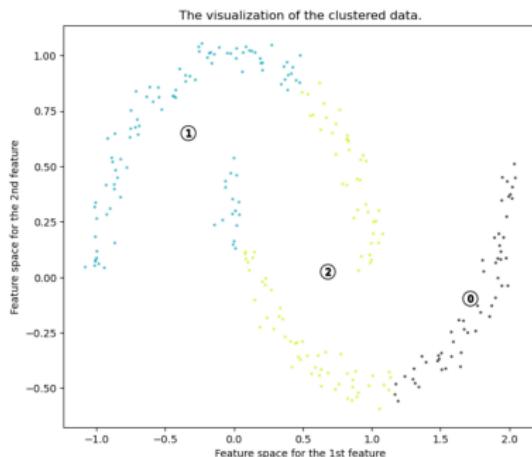
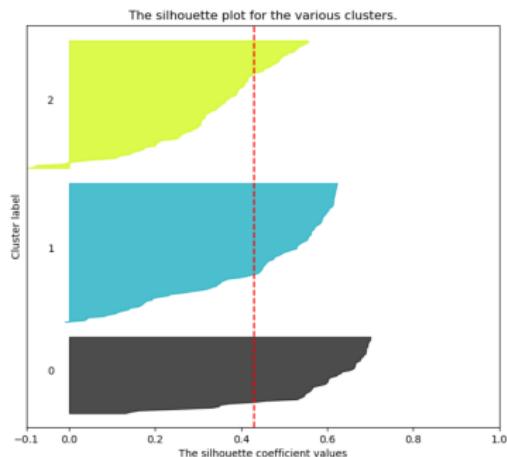


For n_clusters = 2 The average silhouette_score is : 0.49566475412015537

Medidas internas

Visualização da silhueta sobre dados gerados por make-moons

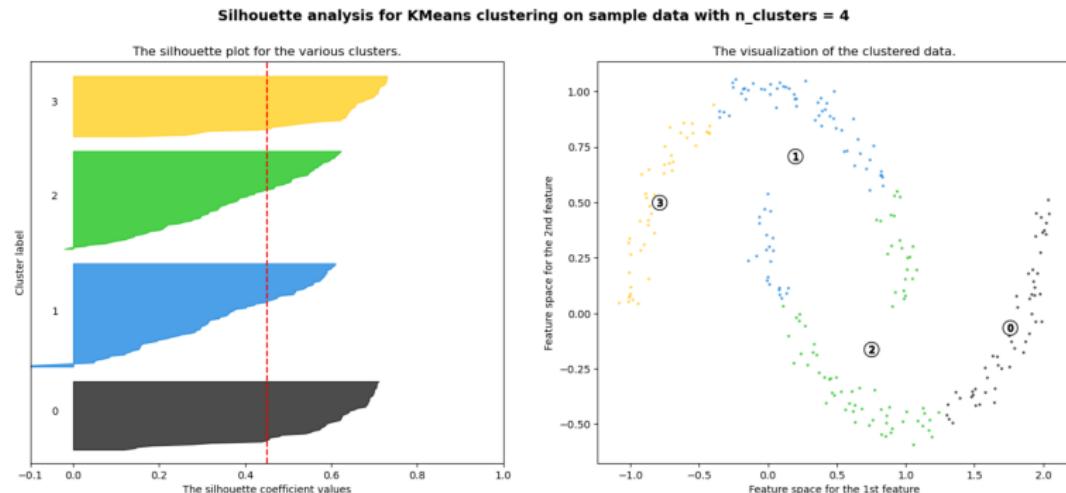
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



For n_clusters = 3 The average silhouette_score is : 0.4291926500333709

Medidas internas

Visualização da silhueta sobre dados gerados por make-moons



For n_clusters = 4 The average silhouette_score is : 0.45008729397689823

Validação de clusters
oooooooo

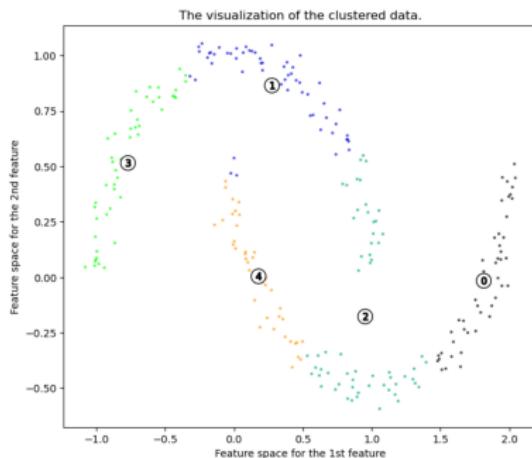
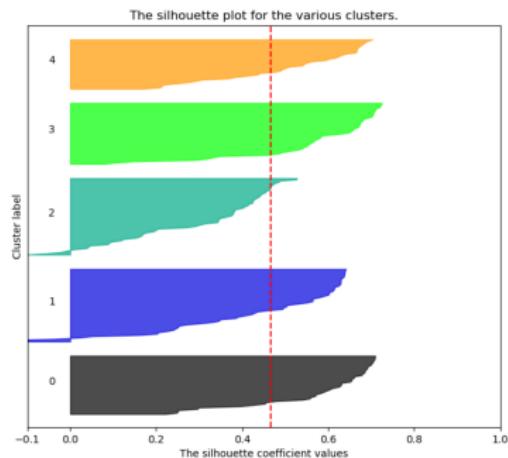
Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooooooooooooooo●oooooooooooo

Medidas internas

Visualização da silhueta sobre dados gerados por make-moons

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



For n_clusters = 5 The average silhouette_score is : 0.4669616060202411

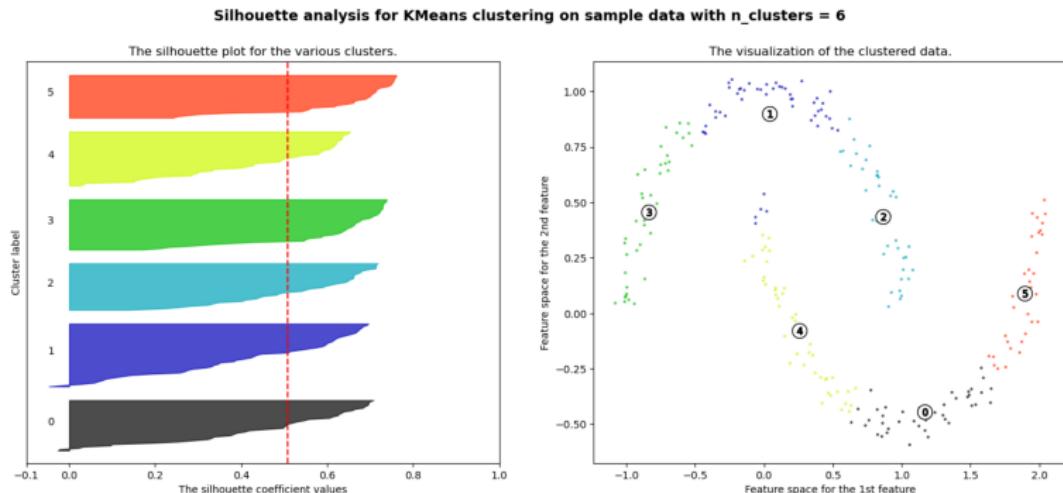
Validação de clusters
oooooooo

Medidas externas
oooooooooooooooooooo

Medidas internas
oooooooooooooooooooo●oooooooooooo

Medidas internas

Visualização da silhueta sobre dados gerados por make-moons



Medidas internas

Método do cotovelo (Elbow method)

- Ideal para determinar o número de centróides para algoritmos que precisam receber como parâmetro o número pré-determinado de clusters a serem buscados.
- O método calcula a soma dos quadrados entre os centróides e cada ponto (SQ).
- Itera-se o algoritmo (ex: k-means) de $k = 2$ até $k = n$ e traça-se um gráfico com os resultados de cada k com SQ .
- Escolhe-se o valor de k no gráfico no ponto onde a curva começa a parecer com uma linha reta.

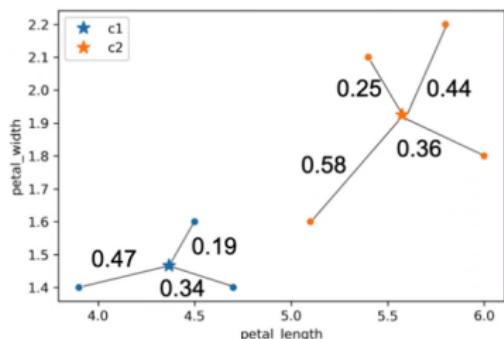
Medidas internas

Método do cotovelo (Elbow method)

- Consideram-se duas medidas para o cálculo do método do cotovelo: a **distorção** e a **inércia**.
- A distorção mede o valor médio das distâncias quadradas dos centros dos respectivos clusters até cada ponto de dados (ex: dist. euclidiana).
 - $Distorcao = \sum_1^k \left(\frac{1}{|c_k|} \sum_1^{|c_k|} dist(ponto, centroide)^2 \right)$
- A inércia mede quão bem um conjunto de dados foi agrupado por K-Means pela soma do quadrado das distâncias dos pontos ao centro mais próximo.
 - $Inercia = \sum_1^k \sum_1^{|c_k|} dist(ponto, centroide)^2$

Medidas internas

Exemplo do Método do cotovelo



Example:

- Inertia: $0.47^2 + 0.19^2 + 0.34^2 + 0.25^2 + 0.58^2 + 0.36^2 + 0.44^2$
- Distortion: $(0.47^2 + 0.19^2 + 0.34^2)/3 + (0.25^2 + 0.58^2 + 0.36^2 + 0.44^2)/4$

Fonte: 23.3 Minimizing Inertia – Data 100, Fall 2020
<https://www.youtube.com/watch?v=YQ2wF0nkelg>

Medidas internas

Método do cotovelo em python

```
from sklearn import metrics
from scipy.spatial.distance import cdist
import numpy as np
import matplotlib.pyplot as plt

distortions = []
inertias = []
mapping1 = {}
mapping2 = {}
K = range(1, 10)

for k in K:
    kmeanModel = KMeans(n_clusters=k).fit(Z)
    kmeanModel.fit(Z)

    distortions.append(sum(np.min(cdist(Z, kmeanModel.cluster_centers_,
                                         'euclidean'), axis=1)) / Z.shape[0])
    inertias.append(kmeanModel.inertia_)

    mapping1[k] = sum(np.min(cdist(Z, kmeanModel.cluster_centers_,
                                         'euclidean'), axis=1)) / Z.shape[0]
    mapping2[k] = kmeanModel.inertia_
```

Medidas internas

Método do cotovelo em python

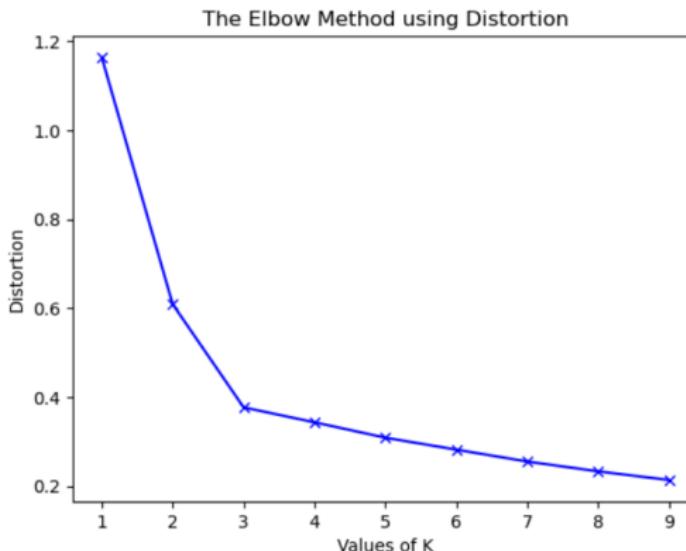
```
for key, val in mapping1.items():
    print(f'{key} : {val}')
```

```
1 : 1.16367658023515
2 : 0.6085815934180989
3 : 0.37669765472257927
4 : 0.34314871377340517
5 : 0.3085086434513629
6 : 0.28164594684038596
7 : 0.2549397483304103
8 : 0.2328265316122258
9 : 0.21336880482461387
```

Medidas internas

Método do cotovelo em python

```
plt.plot(K, distortions, 'bx-')
plt.xlabel('Values of K')
plt.ylabel('Distortion')
plt.title('The Elbow Method using Distortion')
plt.show()
```



Medidas internas

Método do cotovelo em python

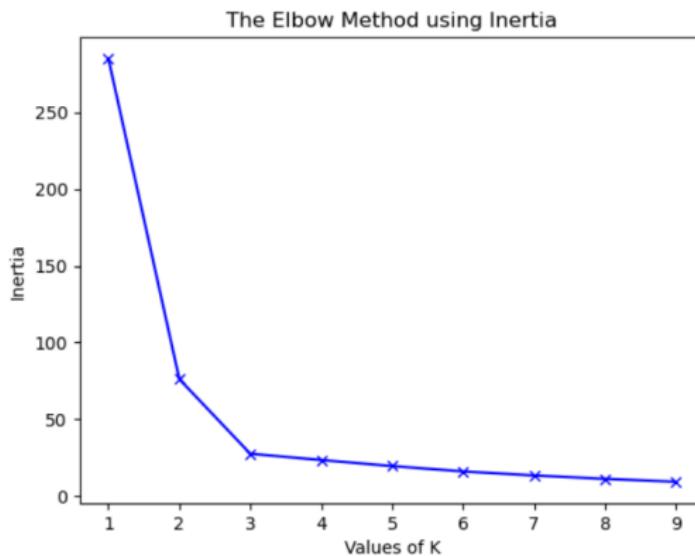
```
for key, val in mapping2.items():
    print(f'{key} : {val}')
```

```
1 : 285.0288422555268
2 : 76.18025819186056
3 : 27.6273703493976
4 : 23.534567201880755
5 : 19.60086157678036
6 : 16.097334206024016
7 : 13.491386824463216
8 : 11.272593300788746
9 : 9.359199008595937
```

Medidas internas

Método do cotovelo em python

```
plt.plot(K, inertias, 'bx-')
plt.xlabel('Values of K')
plt.ylabel('Inertia')
plt.title('The Elbow Method using Inertia')
plt.show()
```



Medidas internas

Índice de Dunn

- Identifica conjuntos de clusters compactos e com pequena variação entre os membros do cluster.
- As médias dos diferentes clusters devem estar suficientemente distantes, em comparação com a variância do cluster.
- Um índice mais alto indica clusters compactos e bem separados, enquanto um índice mais baixo indica clusters menos compactos ou menos bem separados.
- Faixa: {0, 1}
- O custo computacional aumenta com o aumento da dimensionalidade.

Medidas internas

Índice de Dunn

- ① Para cada cluster, calcule a distância entre cada um dos objetos no cluster e os objetos nos outros clusters.
- ② Calcule o mínimo desta distância entre pares como a separação entre clusters ($\delta(C_i, C_j)$).
- ③ Para cada cluster, calcule a distância entre os objetos no mesmo cluster (intra-cluster) (Δ_m).
- ④ Use a distância máxima intra-cluster (ou seja, diâmetro máximo) como a compactação.
- ⑤ Calcule o índice de Dunn.

$$DI_k = \frac{\min_{1 \leq i \leq j \leq k} \delta(C_i, C_j)}{\max_{1 \leq m \leq k} \Delta_m}$$

onde i e j são dois clusters quaisquer, k é o total de clusters e m é um dos clusters resultantes.

REFERÊNCIAS

-  AGGARWAL, C. C.; REDDY, C. K. *Data Clustering: Algorithms and Applications*. 1st. ed. [S.I.]: Chapman & Hall/CRC, 2013. ISBN 1466558210.
-  LIANG, J. et al. Determining the number of clusters using information entropy for mixed data. *Pattern Recognition*, v. 45, p. 2251–2265, 06 2012.
-  MANNHEIMER, J. et al. A systematic analysis of genomics-based modeling approaches for prediction of drug response to cytotoxic chemotherapies. *BMC Medical Genomics*, v. 12, 06 2019.
-  MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. Disponível em: <<http://nlp.stanford.edu/IR-book/information-retrieval-book.html>>. ISBN 978-0-521-86571-5.