



APRENDIZADO EM MÁQUINAS — 1C

Machine Learning — 1C

Prof. Dr. Waldemar Bonventi Jr.

mar2020



PROBLEMINHA...

Como resolver qual o melhor valor de k ,

Ou seja

Qual a melhor partição que “separa bem” os dados?

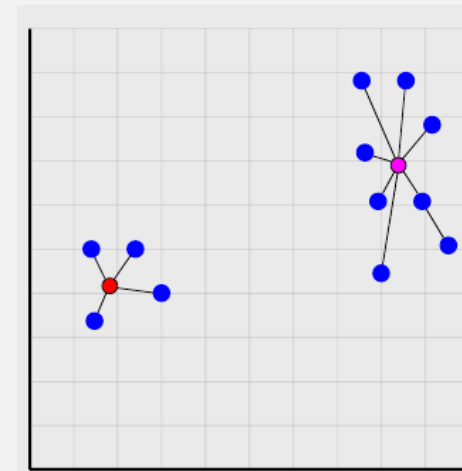
CrITÉrios:

Mínima distância intracluster

Máxima distância intercluster

automovel	peso(kg)	potencia(cv)	bagagem(litros)	Grupo
A	1250	110	650	g2
B	800	80	300	g1
C	900	90	450	g4
D	750	100	400	g4
E	1100	90	350	g1
F	1050	90	600	g3
G	750	70	500	g3

4 grupos?



CRITÉRIO DE QUALIDADE DAS PARTIÇÕES

Pode-se demonstrar que o algoritmo minimiza a seguinte função objetivo (variâncias intra-cluster):

$$J = \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} d(\mathbf{x}_j, \bar{\mathbf{x}}_i)^2$$

onde $\bar{\mathbf{x}}_i$ é o centróide do i -ésimo cluster

Para cada cluster (grupo) formado, calcula-se a distância do elemento x_j ao centro do grupo ao qual pertence \bar{x}_i (somatório “interno”)

Depois, somam-se estas “distâncias somadas” para todos os grupos (clusters) – somatório “externo”

Pode-se demonstrar que o algoritmo minimiza a seguinte função objetivo (variâncias intra-cluster):

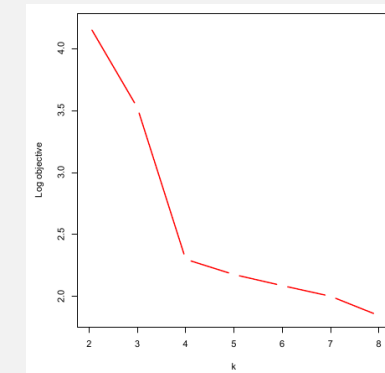
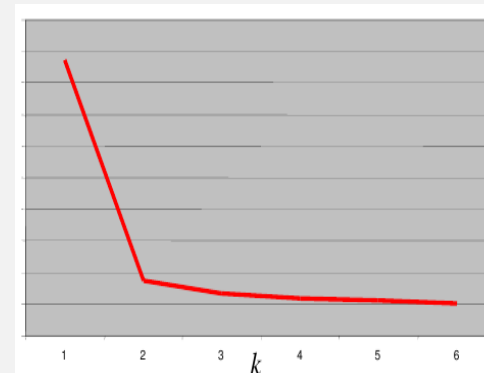
onde \bar{x}_i é o centróide do i -ésimo cluster

Para cada cluster (grupo) formado, calcula-se a distância do elemento x_j ao centro do grupo ao qual pertence \bar{x}_i (somatório “interno”)

Depois, somam-se estas “distâncias somadas” para todos os grupos (clusters) – somatório “externo”

Calculando-se J para vários valores de k

- Isto é, para cada situação de particionamento: $k=2$ grupos, $k=3$ grupos, $k=4$, ...
 - ($k=N$ um grupo por elemento)



MÉTODO COTOVELO (*ELBOW METHOD*)

Aplicar o algoritmo k-means para diferentes valores de k .

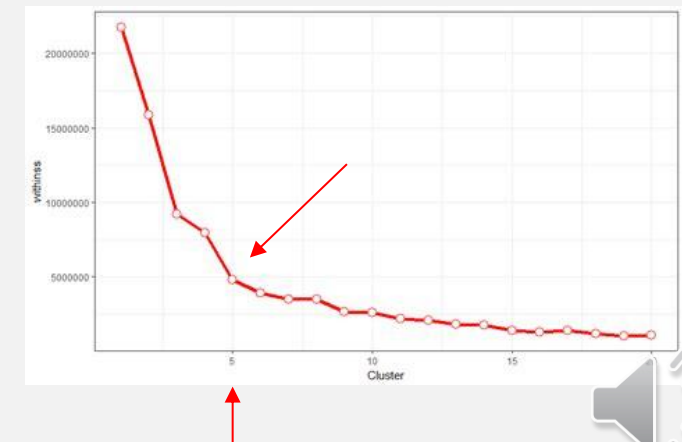
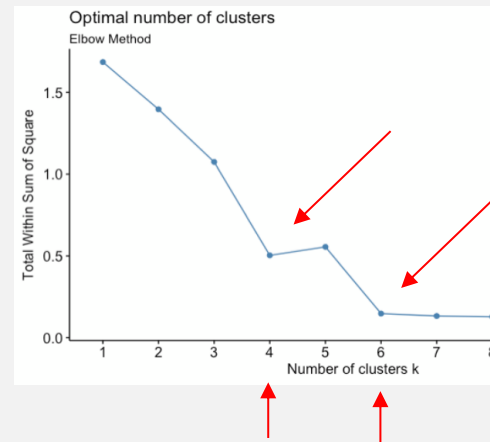
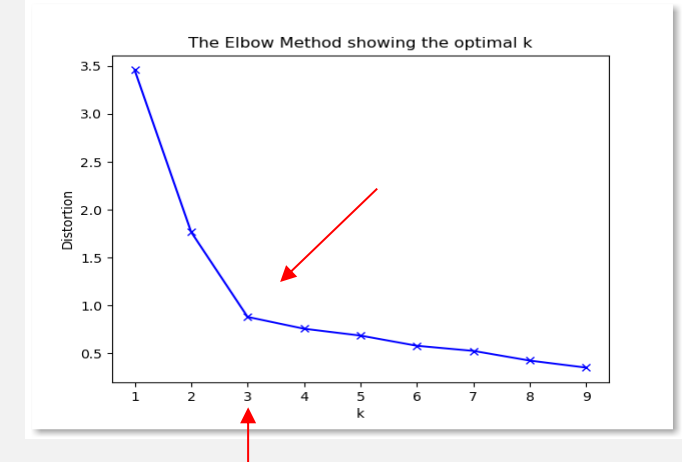
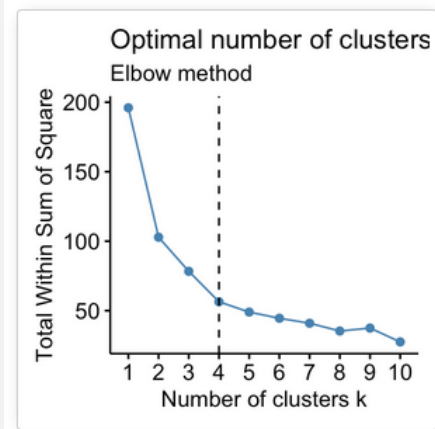
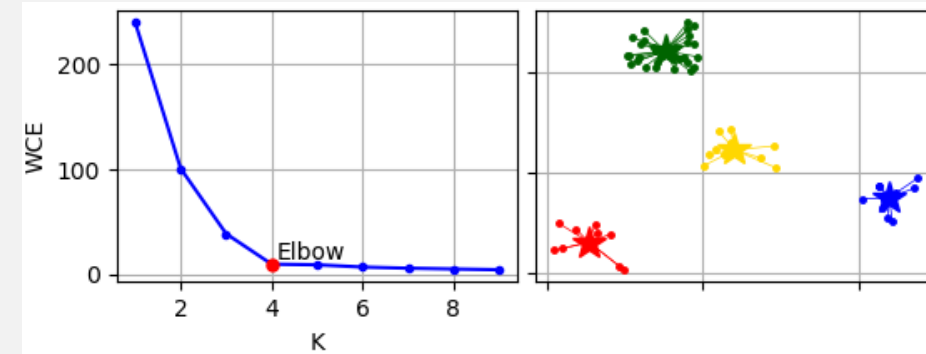
Por exemplo, variando k de 1 a 10 clusters.

Para cada k , calcule a soma total quadrada (J) intracluster.

Plote a curva de J de acordo com o número de clusters k .

A localização de uma dobra (joelho) na plotagem é geralmente considerada como um indicador do número apropriado de clusters.

Aplicar o algoritmo k-means com este número e atribuir a cada objeto o grupo designado



OUTRO MÉTODO: COEFICIENTE DE SILHUETA

O coeficiente da silhueta é calculado usando a distância média intra-cluster (a) e a distância média mais próxima do cluster (b) para cada amostra. O coeficiente de silhueta para uma amostra é

$$(b - a) / \text{máximo}(a, b).$$

Para esclarecer, b é a distância entre uma amostra e o cluster mais próximo do qual a amostra não faz parte. Observe que o coeficiente da silhueta é definido apenas se o número de etiquetas for

$$2 \leq n_{\text{labels}} \leq n_{\text{samples}} - 1.$$

Esta função retorna o coeficiente médio da silhueta em todas as amostras.

O melhor valor é 1 e o pior valor é -1.

Valores próximos a 0 indicam clusters sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao cluster errado, pois um cluster diferente é mais semelhante.

Veja:

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

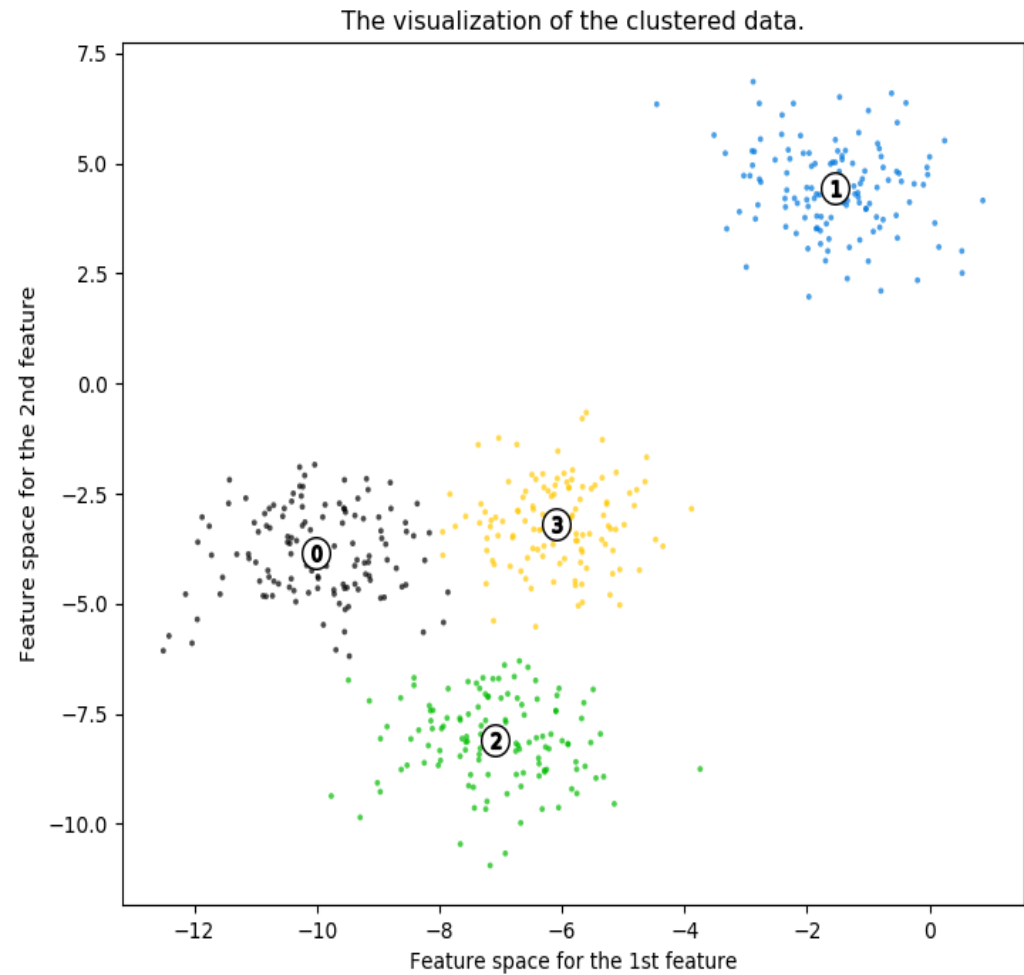
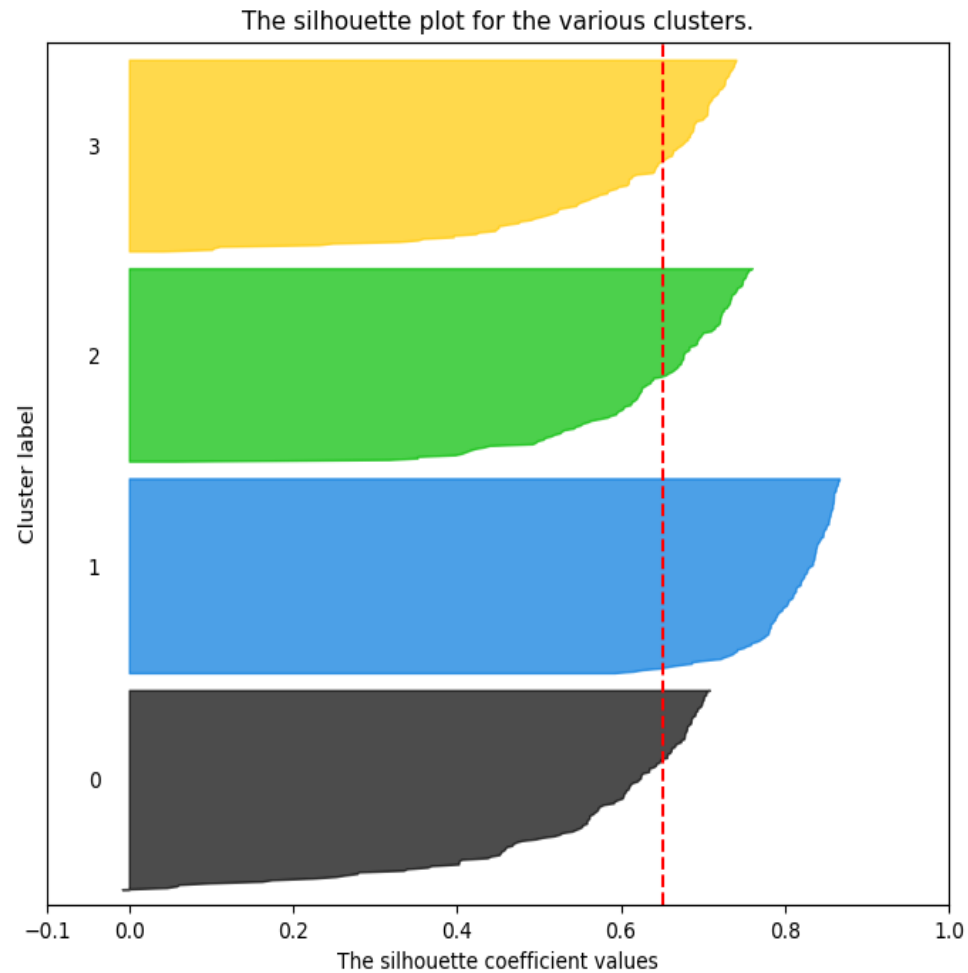
Neste exemplo, a análise de silhueta é usada para escolher um valor ideal para `n_clusters`. O gráfico da silhueta mostra que o valor `n_clusters` de 3, 5 e 6 é uma má escolha para os dados fornecidos devido à presença de clusters com pontuações abaixo da média da silhueta e também devido a grandes flutuações no tamanho dos gráficos da silhueta. A análise de silhueta é mais ambivalente na decisão entre 2 e 4.

Também a partir da espessura do gráfico da silhueta, o tamanho do cluster pode ser visualizado. A plotagem da silhueta para o cluster 0, quando `n_clusters` é igual a 2, é maior em tamanho devido ao agrupamento dos 3 subclusters em um grande cluster.

No entanto, quando `n_clusters` é igual a 4, todos os gráficos têm mais ou menos espessuras semelhantes e, portanto, têm tamanhos semelhantes, como também pode ser verificado na dispersão rotulada



Silhouette analysis for KMeans clustering on sample data with $n_clusters = 4$





PRÓXIMO CAPÍTULO: APLICAÇÃO |

