K-Nearest Neighbor (KNN): o que está por trás do software?

[1]Edilton de Souza Barcellos[1*+] (PQ), Márcia Miguel Castro Ferreira[2] (PQ).

[1]Departamento de Química/UFV, [2] Instituto de Química/Unicamp – [1]barcello@ufv.br*

Palavras Chave: KNN, Análise de Agrupamentos, single linkage, similaridade, medida de distâncias.

Introdução

A compreensão dos processos ocorrendo no mundo requer o uso de diferentes metodologias em concordância com os objetivos em vista. Em muitas áreas de estudo é necessário agrupar objetos e/ou variáveis que possuam características similares.

Um método bastante usado para agrupar objetos em função de similaridades e distâncias é o KNN. Pode ser usado em vários campos tais como psicologia, biologia, química, ciência de alimentos medicina, ciência ambiental dentre outros[1].

Esse trabalho apresenta um passo a passo envolvendo cálculos de distâncias e similaridades necessários à obtenção dos resultados na forma de um dendrograma.

Parte-se de um conjunto de dados simulados contendo sete objetos (amostras) e duas variáveis.

Resultados e Discussão

A matriz X(7x2) dada abaixo, pode representar as concentrações de dois analitos (as variáveis) nas sete amostras coletadas em pontos diferentes de um dado local (lago, rio, solo, etc); em alimentos (marcas diferentes de um produto); dentre outros.

X=[1 10;13 12;4 8;3 6.5;10 5;10 3.5;16 4.5];

A Fig.1 representa um gráfico da variável 1 versus variável 2 da matriz de dados originais, destacando a junção dos pontos (5 e 6), como se vê abaixo.

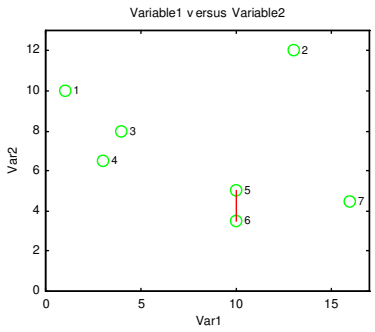

Figura 1.. Junção dos pontos 5 e 6, os quais possuem menor distância entre si.

O resultado do 1º passo, usando agrupamento por conexão simples ("single linkage"), gera a matriz de distâncias, D1, adiante.

D1 =

12.1655	3.6056	4.0311	10.2956	11.1018	15.9765
0	9.8489	11.4127	7.6158	9.0139	8.0777
0	0	1.8028	6.7082	7.5000	12.5000
0	0	0	7.1589	7.6158	13.1529
0	0	0	0	1.5000	6.0208
0	0	0	0	0	6.0828

Calculam-se distâncias para a construção das matrizes D2, D3, D4, e assim em diante até que todos os pontos estejam ligados.

Com base nas expressões matemáticas desses cálculos[2] (omitidas aqui) chega-se ao Dendrograma conforme se vê na Figura abaixo.

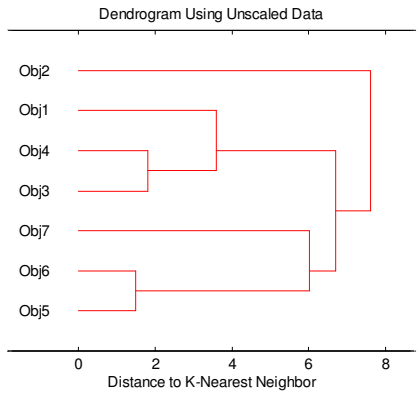

Figura 2. Dendrograma correspondente ao agrupamento ("single linkage") de dados alocados em uma matriz de 7 objetos e duas variáveis.

Comparando-se os pontos 5 e 6 em cada figura vê-se claramente que eles são os mais próximos. Além disso, na Fig.2, vê-se na escala horizontal que esses dois pontos possuem o menor valor (~1,5).

Conclusões

Vê-se, nas Fig. 1 e 2 uma perfeita concordância para todos os pontos. Tanto as matrizes quanto os gráficos obtidos a partir delas são simples. A construção do dendrograma se dá de forma natural.

Agradecimentos

Aos Professores Luiz Henrique Mendes da Silva e Maria do Carmo Hespanhol da Silva (DEQ/UFV), eternos incentivadores.

[1] Everitt, B. , Heinemamm Educational Books ltd, Social Sience Research Council, London, 1974.
[2] Sharaf, M. A.; Illman, D. L.e Kowalski, B. R. John Wiley & Sons, 1986.