04

Matemática do Calinski (Parte 1)

Transcrição

Neste vídeo, conheceremos o índice Calinski-Harabasz, e ele leva em consideração a dispersão dos pontos dentro de um cluster, tanto nos pontos dentro de um cluster específico, como entre clusters diferentes.

Trata-se de uma fórmula mais complicada que as outras, divida em duas partes principais. A primeira, nos fornece a razão entre a dispersão dentro do cluster e entre clusters. Já a segunda parte, o valor é multiplicado em relação ao número de clusters e elementos.

Começaremos por resolver a segunda parte da fórmula.

Índice Calinski-Harabasz

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

O primeiro valor que temos é de Ne, isto é, número de elementos dentro do cluster. Temos três clusters e três elementos em cada um deles, logo 9 elementos no total.

K, é o número de klusters, então, 3.

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{9 - 3}{3 - 1}$$

Feitas as substituições, estamos prontos para finalizar essa parte da oração. O resultado será 3.

Já podemos começar a trabalhar com a primeira parte da fórmula. Terremos dois valores : Bk, dispersão dos elementos entre clusters e Wk, dispersão de elementos dentro do cluster.

Começaremos com o valor de Wk, que demanda uma nova equação. A primeira parte dessa nova equação nada mais é que um somatório, somaremos o valor que está à direita para cada um dos clusters.

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

k = número de clusters
q = cluster

Já a segunda parte consiste estimar uma matriz de variância-covariância para cada uma das viaríaveis que temos, isto é, nossos atributos. Montaremos a nossa matriz passo a passo.

A primeira etapa é obter uma matriz com os valores de x e y. Começaremos com o cluster verde. Já obtemos esses valores anteriormente, assim como o vetor com os centróides deste cluster. Com base nestes valores, obteremos uma nova matriz com as diferenças em relação aos centróides.

Essa nova matriz conterá os valores de x' e y'. Para calcular os valores de x', diminuiremos os valores de x pelo centróide 1,1. O mesmo será feito para y.

Matriz de Variância-Covariância

3) Obter uma matriz com as diferenças em relação aos centróides

👨	x'	y'
A	1,0 - 1,1	0,9 - 1,3
B	1,0 - 1,1	1,7 - 1,3
C	1,3 - 1,1	1,5 - 1,3

👨	x	y
A	1,0	0,9
B	1,0	1,7
C	1,3	1,5

👨	x	y
ct	1,1	1,3

Ao obtermos os valores valores corretamente, criaremos uma nova matriz transposta, isto é, uma matriz inversa. Uma coluna se transforma em linha e uma linha se transforma em coluna.

4) Criar uma matriz transposta

👨	x'	y'
A	-0,10	-0,40
B	-0,10	0,40
C	0,20	0,20

👨	A	B	C
x'	-0,10	-0,10	0,20
y'	-0,40	0,40	0,20

Multiplicaremos essas duas matrizes, e então teremos o resultado de uma terceira, a matriz final de variância-covariância.

👨	x'	y'
x'	var (x', x')	cov (x', y')
y'	cov (y', x')	var (y', y')

Calcularemos os valores de variância para x, y e entre x e y e y e x. Coletaremos cada um desses valores. O valor da variância de x será de 0,06, e a variância de y será de 0,36.

Não calcularemos os valores de covariância pois ela não será utilizada no cálculo, ela está aqui agora apenas para nos auxiliar a criar a matriz.