06

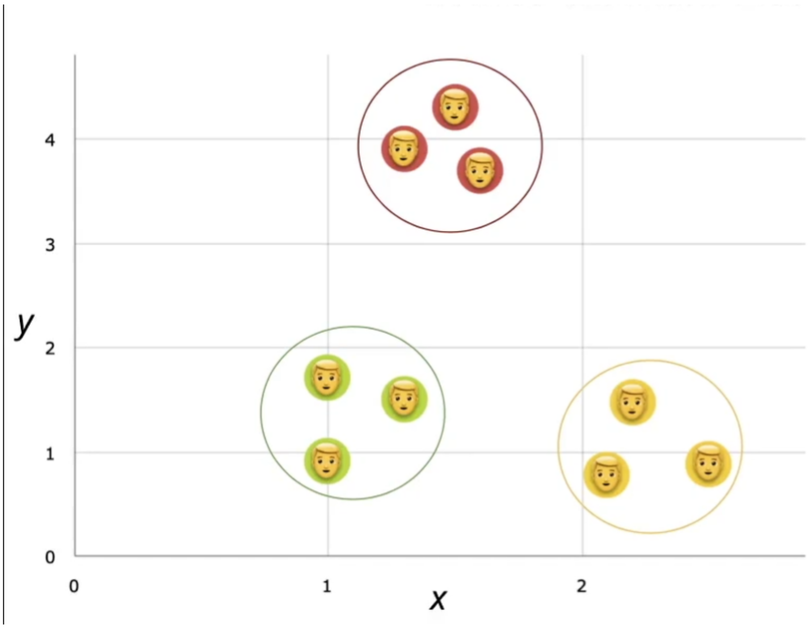
Matemática do Calinski (Parte 3)

Transcrição

Temos a fórmula de Bk, que também se estrutura em duas partes. Na segunda parte, temos alguns valores diferentes, como Nq.

$$B_k = \sum_{q=1}^{k} \boxed{n_q(c_q - c_E)(c_q - c_E)^T}$$

Nq nada mas é que o número de elementos no cluster, enquanto Cq se refere ao centroide do cluster e Ce o centroide dos elementos. Já temos alguns dos valores solicitados na fórmula, como Cq e Nq.



	x	y
ct	1,1	1,3

	x	y
ct	1,4	3,9

	x	y
ct	2,2	1,0

Contudo, ainda não temos o resultado de Ce. Para isso, coletaremos todos os pontos e tiraremos a média para cada uma das variáveis.

	x	y
A	1,0	0,9
B	1,0	1,7
C	1,3	1,5
D	1,3	3,9
E	1,5	4,3
F	1,6	3,7
G	2,1	0,8
H	2,2	1,5
I	2,5	0,9

Índice Calinski-Harabasz

	x	y
ce	1,61	

Feito isso, mediremos a distância entre o centroide de um cluster com o centroide total.

	x'	y'
ct	-0,51	-0,83

	x	y
ct	1,1	1,3

	x	y
ce	1,61	2,13

Elaboraremos a matriz transposta, e coletaremos primeiramente a variância de x. Basicamente multiplicaremos uma cédula só, porque temos menos dados. Multiplicaremos o valor de -0,51 por ele mesmo, e teremos 0,2 par x. No caso de y,o resultado será -0,83 multiplicado por ele mesmo e teremos 0,68.

Multiplicaremos esses dois valores pelo número de elementos que existem dentro do cluster.

	x'	y'
x'	0,78	cov (x', y')
y'	cov (y', x')	2,04

Faremos exatamente o mesmo procedimento para os clusters vermelho e amarelo.

O último passo é somar as matrizes de variância geradas.

Soma das Matrizes

	x'	y'
x'	1,92	-
y'	-	15,24

No resultado, conseguiremos ver que o valor de y é muito maior que o de x. A discussão se dá em termos internos de um cluster, e aqui queremos aumentar a dispersão entre elementos, portanto possui um bom valor.

Somaremos os dois valores para obter o traço de Bk, que será 17,16. Substituiremos esse valor na fórmula

$$s = \frac{17{,}16}{1{,}08} \times 3$$

O resultado será 47,64. Esse é o valor final para nosso índice Calinski-Harabasz.