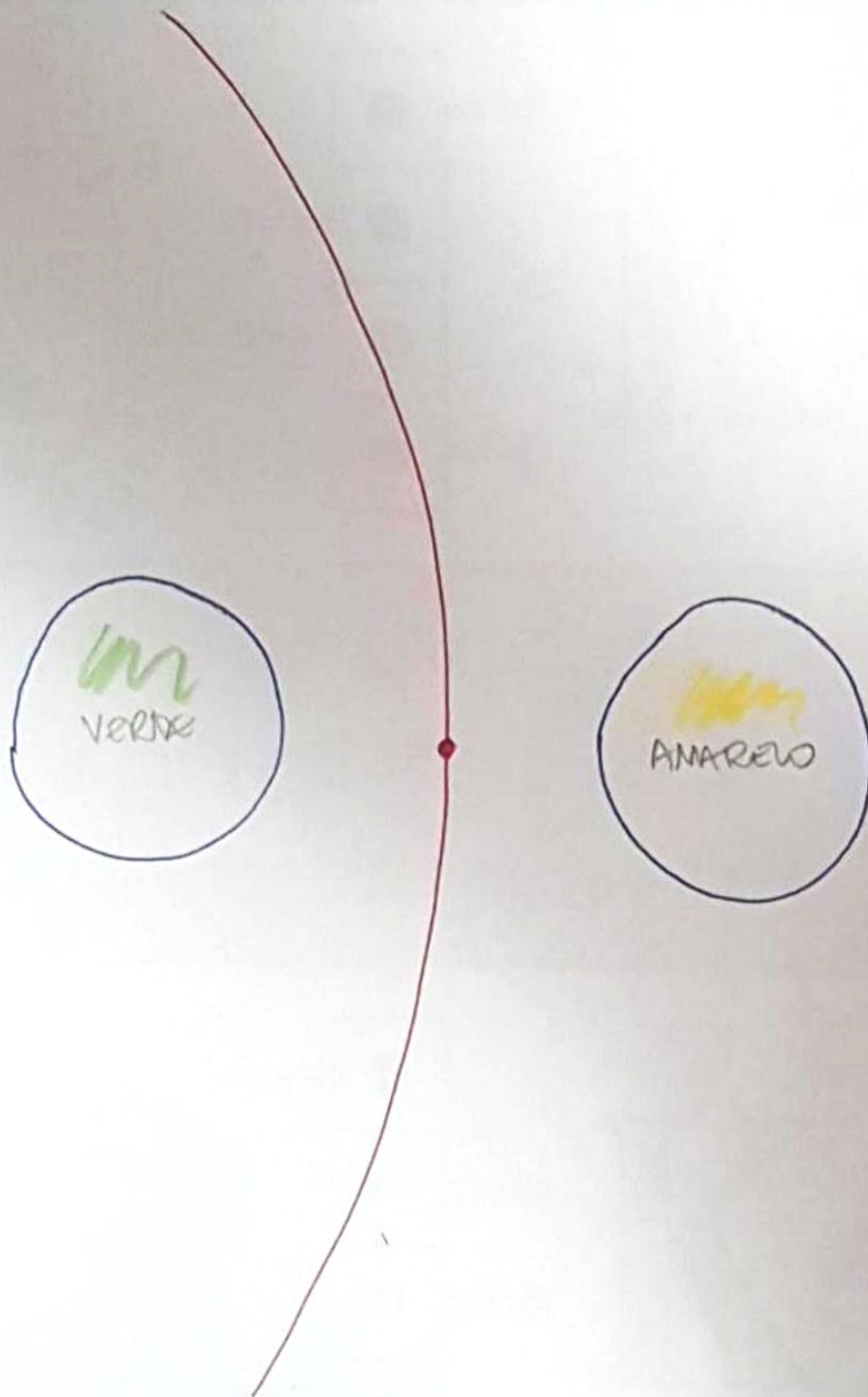


KMEANS



TRABALHO DEDICADO A

MARY FAIRFAX SOMERVILLE
(1780 - 1872)

1.1 KMEANS

1

1.1 KmeanS Clustering Method

1

Particionamento de n dados
d-dimensionais (X_1, X_2, \dots, X_n) :

$$\begin{array}{ccc}
 d = 1 & d = 2 & d = 3 \\
 \left| \begin{array}{c} (X_{11}) \\ (X_{11}) \\ \circ \circ \circ \\ (X_{11}) \end{array} \right| & \left| \begin{array}{c} (X_{11}, X_{12}) \\ (X_{21}, X_{22}) \\ \circ \circ \circ \\ (X_{n1}, X_{n2}) \end{array} \right| & \left| \begin{array}{c} (X_{11}, X_{12}, X_{13}) \\ (X_{21}, X_{22}, X_{23}) \\ \circ \circ \circ \\ (X_{n1}, X_{n2}, X_{n3}) \end{array} \right| \\
 \leftarrow d = d \rightarrow \\
 \left| \begin{array}{c} (X_{11}, X_{12}, X_{13}, X_{14}, \dots, X_{1d}) \\ (X_{21}, X_{22}, X_{23}, X_{24}, \dots, X_{2d}) \\ \circ \circ \circ \\ (X_{n1}, X_{n2}, X_{n3}, X_{n4}, \dots, X_{nd}) \end{array} \right|
 \end{array}$$

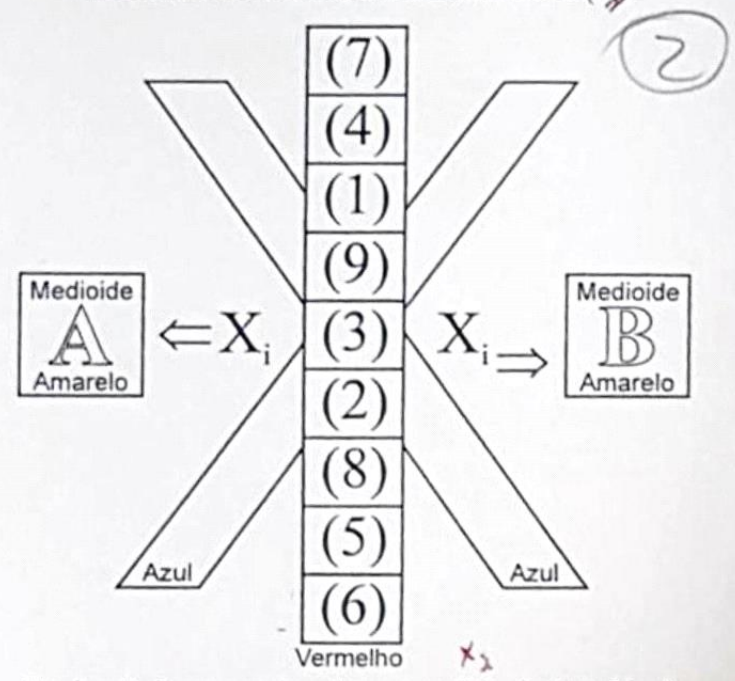
Em K clusters ou K politopos ($K \leq n$) nucleados por k medioides minimizadores das distâncias entre esses medioides e os dados formadores dos seus respectivos clusters.

Politopos de Veronoi

3.0 Construção Epistêmica do Método de MacQueen

Kmeans Monodimensional 2means

Particionar nove dados monodimensionais (X_i):



Em $k = 2$ clusters, cujos elementos (X_i), são "atraídos" por dois medióides.

Within Sum

KMEANS

$$d_1 = \sqrt{(x_2 - x_1)^2}$$

$$d_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

... A ESTABILIDADE DOS MEDIOIDES OBTIDOS COM AS SEMENTES SORTEADAS PODE SER MELHORADA COMPUTACIONALMENTE FAZENDO ...

$$d_3 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

$$d_4 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \dots + (w_2 - w_1)^2}$$

WITHIN SUM

3

5.0 Computação do KmeanS

3

Para um big Data particionado em K clusters ($K = 1, 2, 3, \dots$), de respectivos medioides (\bar{m}_k), clusters formados por e_{ki} elementos ($i = 1, 2, 3, \dots$); computa-se sua *Within Sum* (WS):

$$WS = \sum_K^{\text{Clusters}} \sum_i^{\text{Elementos}} [e_{ki} - \bar{m}_K]^2$$

Within Sum

Amarelo

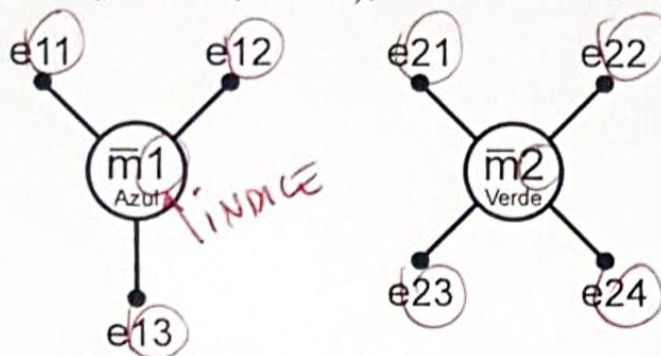
O que permite formatar a praxis matemática do KmeanS:

Big Data Within Sum Application

(4)

5.1 Exemplo do Cálculo do WS

Para $K = 3$, três medioides (\bar{m}_1 , \bar{m}_2 , \bar{m}_3) e 9 elementos, sendo $3 + 4 + 2$, respectivamente os elementos do clusters ($K=1$, $K=2$, $K=3$), resulta a semiose:



$$WS = \sum_{i=1}^3 [e_{1i} - \bar{m}_1]^2 + \sum_{i=1}^4 [e_{2i} - \bar{m}_2]^2 + \sum_{i=1}^2 [e_{3i} - \bar{m}_3]^2$$

Within Sum



Within Sum

(A)

Big (A)

INÍCIO

Por ser necessário os elementos e_{ki} do Big Data estarem o mais avizinjado possível do seu medioide \bar{m}_k , muda-se esses elementos e_{ki} um de cada vez, do seu cluster original para um outro cluster sorteado e avalia-se novamente o WS.

NÃO
TEM
ESTA
NECESSIDADE

Data

3

6.0 Cálculo do Elbow Data Chart

O número ideal de clusters k^* , "an ideal number of nearest neighbors", é indicado pelo fenômeno *Elbow Data*, que é resumido na observação do valor estatístico WS, (*Within Sum*):

WS
decreases
as
K
gets larger

3

Cotovelo do Big Data

7.2 Elbow Data Chart

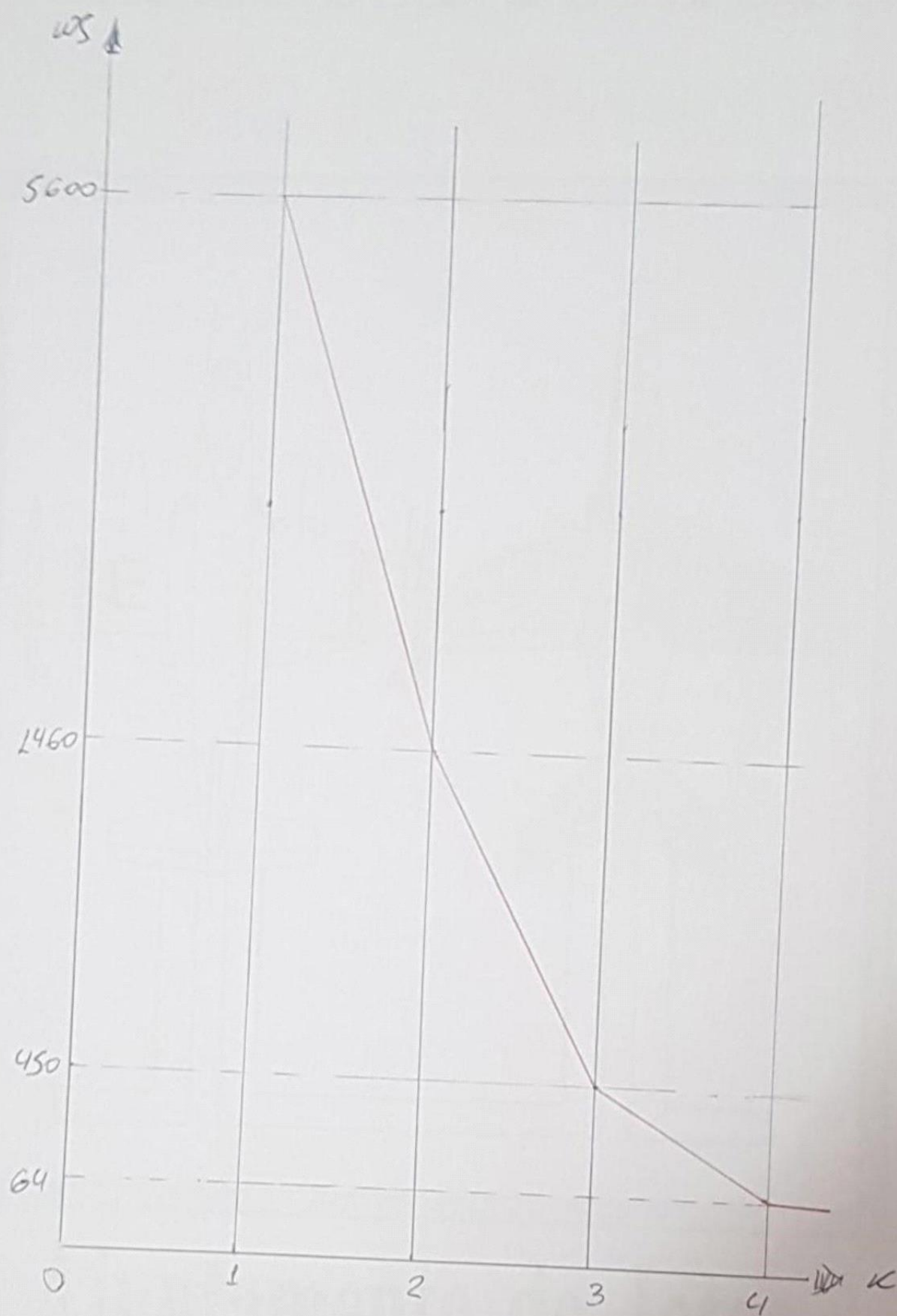


Gráfico Cotovelo

Modelo CivitaS

1.0 Valoração Justa do Tributo IPTU


O Imposto Predial e Territorial Urbano, tributo brasileiro cobrado daqueles que são donos de propriedades imobiliárias

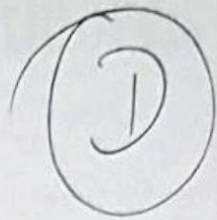
- (Lei nº 5172, 25/10/1966) -

tem como objetivo, além de fiscal, controlar

as valorações e as ocupações das

propriedades brasileiras.

 **meanS** como Equanimidade Social

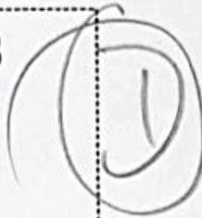


3

3.0 Kmeans Aplicado na Clusterização de Imóveis

Z0 Z1 Z2 Z3 Z4

Sendo 400 imóveis
selecionados
e
classificados com
10 atributos e 40 nuances,
pede-se para particioná-los,
inicialmente,
em
4 grupos homogêneos.



Z5 Z6 Z7 Z8 Z9

Cidade Polita do Sul

AJ

7.2 Observações Pertinentes ao CivitaS

AJ

Primae Observationis

O imóvel 368, no grupo 4, faz o papel de mediodide

0,50	0,50	0	0	0,33	0,33	0,25	0,25	1	0
------	------	---	---	------	------	------	------	---	---

Enquanto o imóvel 140, que também pertence ao grupo 4, é o mais afastado do mediodide ($d^2 \cong 2,80$):

0	1	0,80	0,80	1	1	0	0	1	0
---	---	------	------	---	---	---	---	---	---

Secundo Observationis

Segue a distância entre os imóveis 368 e 140:

$$\begin{aligned} d^2 &= (0,50 - 0)^2 + (0,50 - 1)^2 + (0 - 0,80)^2 + (0 - 0,80)^2 \\ &+ (0,33 - 1)^2 + (0,33 - 1)^2 + (0,25 - 0)^2 + (0,25 - 0)^2 + \\ &+ (1 - 1)^2 + (0 - 0)^2 \end{aligned}$$

$$\begin{aligned} d^2 &= 0,25 + 0,25 + 0,64 + 0,64 + 0,449 + \\ &+ 0,449 + 0,062 + 0,062 + 0 + 0 \cong 2,80 \end{aligned}$$

Modelo Matemático CivitaS

A2

Agências de Rating

A2

Avaliações dos países particionados em 22 clusters:

Conjunto de Dados e Classificações Arbitradas	
PIB PIB per Capta Investimentos Poupança Inflação Importação Exportação Receitas Despesas Dívida Externa Dívida Interna Fluxo de Capitais	(Excelente) AAA AA+ , AA , AA- BBB+ , BBB , BBB- BB+ , BB , BB- B+ , B , B- CC+ CC CC- DDD , DD , D (Inadimplente)
Standart and Poor's: Elbow 22	

Aplicação do KmeanS

A3

Cientista de Dados

A3

A quantidade e a qualidade do conteúdo epistêmico do traço atual, causa-e-efeito das sucessivas alterações de seu aspecto são exegeses autorizadas da sua aparência antecedente, a qual, por sua vez, é eisegese desautorizada da sua aparência subsequente.

Big Data como uma abordagem abducional sígnico-aspectual de possíveis conteúdos matemáticos existentes em Bancos de Dados, submete o Cientista de Dados, sujeito dessa ação epistêmica, após análise desse traço, à dois estados cognitivos não mutualmente exclusivos:

Logográfico e Lexicográfico

A4

SignoMatemA

A4

No **Estado Logográfico** o traço aponta para um encadeamento mínimo de operações lógicas, permitindo alguma interpretação; quando e onde, faz-se necessário contextualizações dessas escassas unidades de sentido menores.

No **Estado Lexicográfico** o traço aponta para um encadeamento máximo de operações resultando em unidades de sentido maiores que, para serem entendidas, dispensam contextualizações.

É, portanto, esse limitado reconhecimento de um possível conteúdo matemático associado ao traço, enquanto marca de descoberta, que torna o Cientista de Dados um paraléxico sígnico, uma vez que esse profissional modifica e troca traços na busca de perceptos.

HomoSignuM