FIAP

MBA+ | Artificial Intelligence & Machine Learning


AI

Aprendizado não supervisionado

K-means / EM / Regras de Associação

Prof. Dr. Vinicius F. Caridá

EM (Expectation Maximization)

Agrupamento – EM (Expectation Maximization)

- O EM é um procedimento geral de aprendizado a partir de dados parcialmente observados.
- Dados um conjunto de variáveis observadas X (X = {X_1, X_2, X_3,X_4}, por exemplo) e um conjunto de variáveis não observadas Z (Z={Y}, por exemplo), o procedimento é:

Enquanto não atingir convergência, faça:

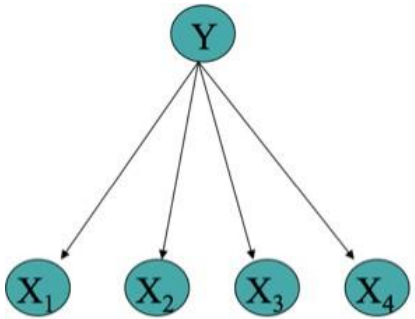
Passo E: com base em X e seus parâmetros θ, calcule P(Y|X, θ);

Passo M: atualize os parâmetros θ, com base nos valores que maximizam as probabilidades encontradas no passo E.

- O Procedimento tem garantia de convergência, mas pode parar em um ótimo local.

EM com base no Classificador Naive Bayes

- Considere que se tenha, como já descrito no slide anterior: um conjunto de variáveis observadas X (X = {X_1, X_2, X_3,X_4}) e um conjunto de variáveis não observadas Z (Z={Y}), o EM tem como base o cáculo da probabilidade P(Y|X,θ). Se assumirmos independência condicional entre todas as variáveis de X e a classe Y, podemos utilizar o Naive-Bayes no procedimento EM.



Y	X1	X2	X3	X4
1	0	0	1	1
0	0	1	0	0
0	0	0	1	0
?	0	1	1	0
?	0	1	0	1

Classificador Naive Bayes

- Exemplo anterior (dados não rotulados sendo usados no treinamento de um classificador Naive Bayes) pode ser visto como um caso de aprendizado semissupervisionado.

- O Algoritmo EM (Expectation-Maximization) funciona como no exemplo anterior

Classificador Naive Bayes

- O EM não é originalmente definido com o Naive Bayes, mas sim com um modelo probabilístico tradicional

- O EM é um algoritmo probabilístico de agrupamento muito comum. No agrupamento, todos os dados de treinamento não possuem rótulo

https://scikit-learn.org/stable/modules/mixture.html#

Validação

"*The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a* ***black art*** *accessible only to those true believers who have experience and great courage.*"

(Jain and Dubes, *Algorithms for Clustering Data*, 1988)

Validação

- Refere-se, de forma ampla, aos diferentes procedimentos para avaliar de maneira objetiva e quantitativa os resultados de análise de agrupamento.
- Cada um desses procedimentos pode nos ajudar a responder uma ou mais questões do tipo:
  - Encontramos grupos de fato?
    - grupos são pouco usuais ou facilmente encontrados ao acaso?
  - Qual a qualidade (relativa ou absoluta) dos grupos encontrados?
  - Qual é o número natural / mais apropriado de grupos?

Índices de validação

A maneira quantitativa para validação é alcançada através de algum tipo de **índice**. Há 3 tipos de **índices/critérios de validade**:

- **Internos**: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados.
- **Relativos**: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa.
- **Externos**: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento esperada ou conhecida.

Índices de validação: Internos

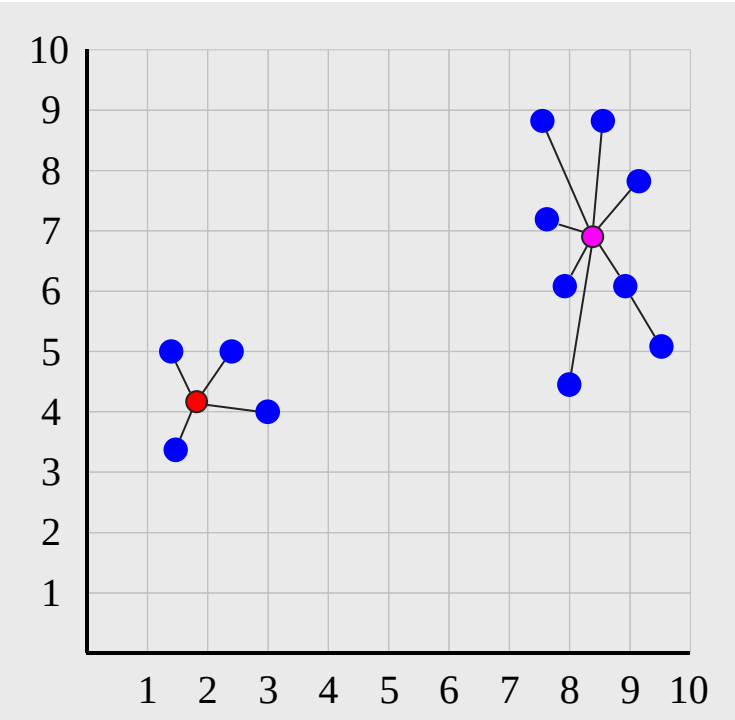
A maneira quantitativa para validação é alcançada através de algum tipo de **índice**. Há 3 tipos de **índices/critérios de validade**:

- **Internos**: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados.
- **Relativos**: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa.
- **Externos**: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento esperada ou conhecida.

Índices de validação: Internos – Erro Quadrático

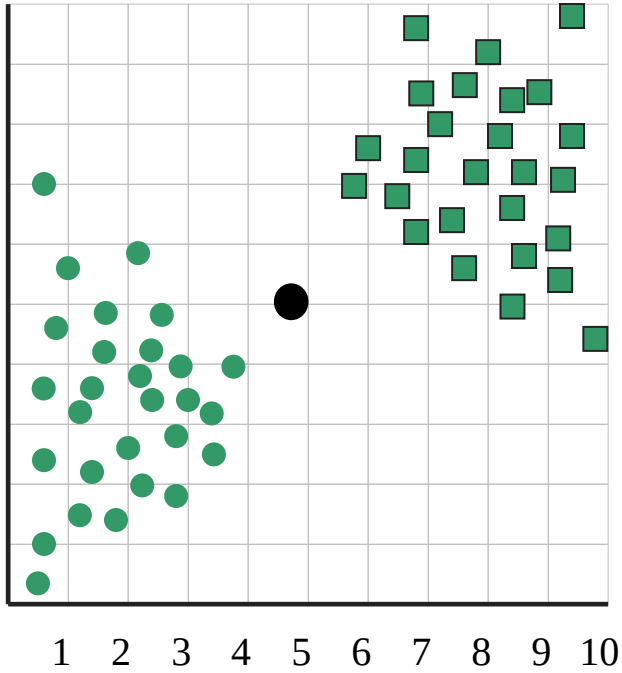
$$J = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_i} d\left(\mathbf{x}_j, \bar{\mathbf{x}}_i\right)^2$$

Função Objetivo



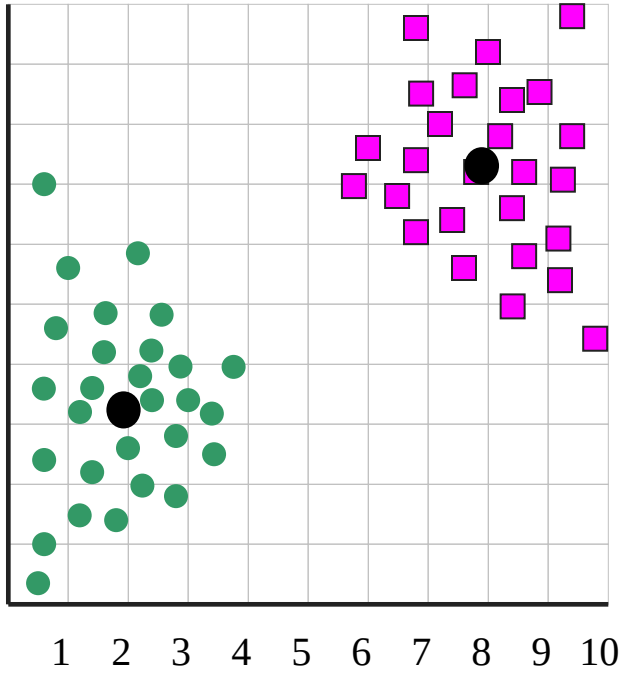
Exemplo de Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Índices de validação: Internos – Erro Quadrático



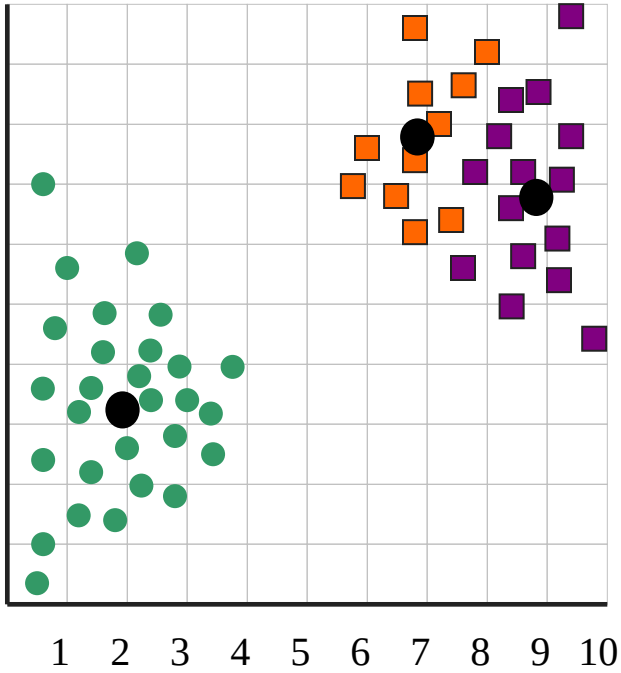
Para $k = 1$, o valor da função objetivo é 873

Índices de validação: Internos – Erro Quadrático



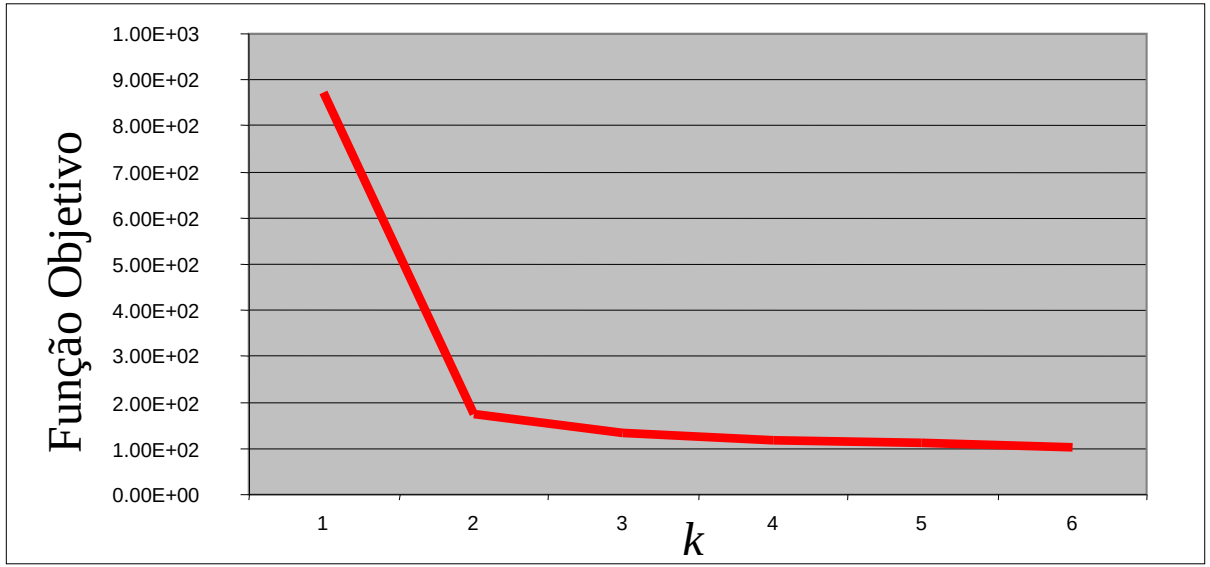
Para $k = 2$, o valor da função objetivo é 173

Índices de validação: Internos – Erro Quadrático



Para $k = 3$, o valor da função objetivo é 134

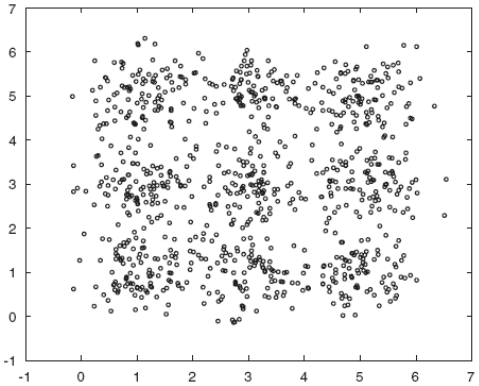
Índices de validação: Internos – Erro Quadrático

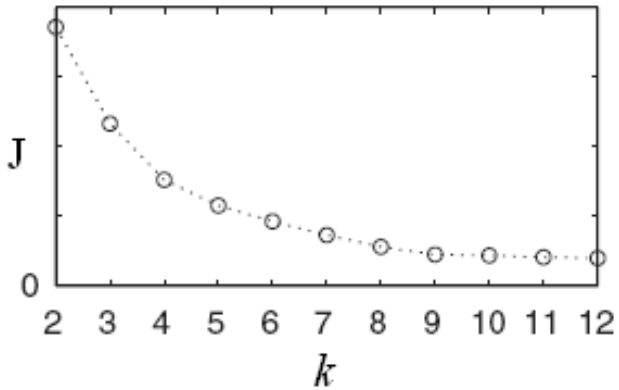


Podemos então repetir este procedimento e plotar os valores da função objetivo *J* para *k* =1,…,6, ... e tentar identificar um “joelho”

Índices de validação: Internos – Erro Quadrático

Infelizmente os resultados não são sempre tão claros quanto no exemplo anterior:





Outras alternativas para lidar com o problema de se estimar o número de clusters?

- Índices de validade relativos...

Índices de validação: Relativo

A maneira quantitativa para validação é alcançada através de algum tipo de **índice**. Há 3 tipos de **índices/critérios de validade**:

- **Internos**: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados.
- **Relativos**: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa.
- **Externos**: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento esperada ou conhecida.

Índices de validação: Relativo - Silhueta

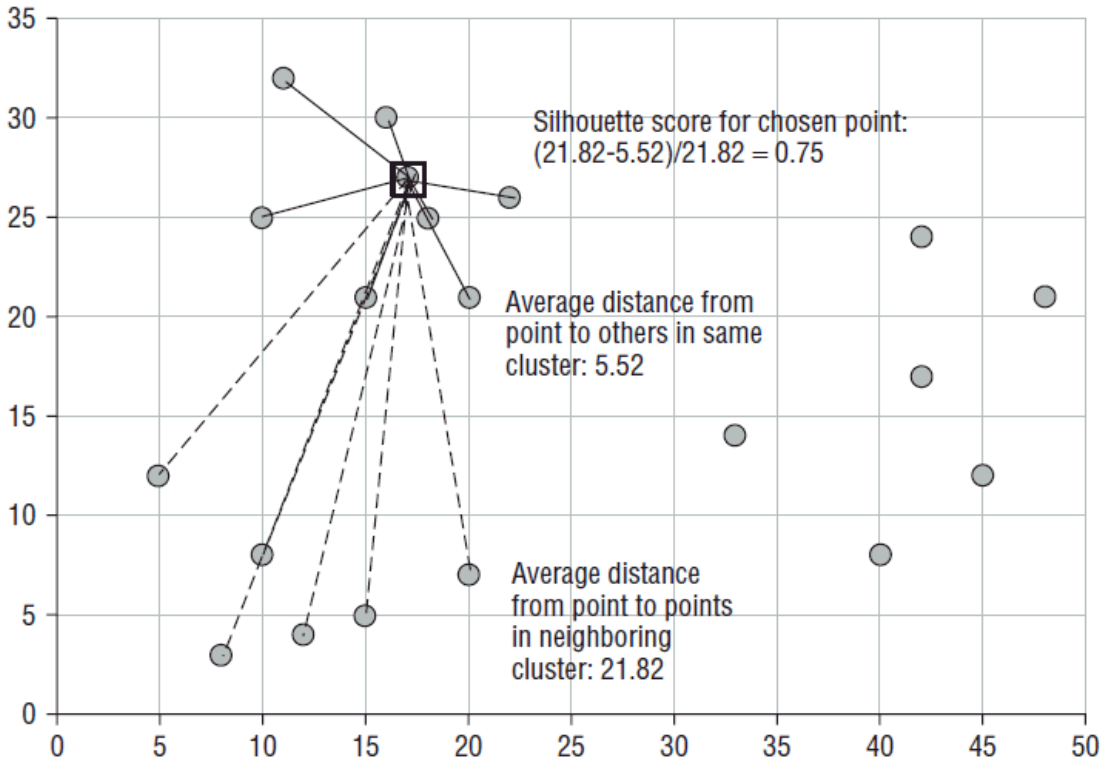
Métrica utilizada para mensurar a qualidade do agrupamento priorizando grupos densos/concisos e separados.

$$s_{x_j} = \frac{b_{p,j} - a_{p,j}}{\max\{b_{p,j}, a_{p,j}\}}$$

Considere que o j-ésimo objeto, x_j, de um data set pertence a um cluster $p \in \{1, \ldots, k\}$

- $a_{p,j}$ - distância media do objeto j a todos os outros objetos pertencentes ao mesmo cluster p
- $d_{q,j}$ - distância media do objeto j a todos os outros objetos pertencentes a clusters distintos q em que $q \neq p$
- $\mathrm{b}_{p,j}$ - o menor valor de $d_{q,j}$ computado para $q = 1, \ldots, k$ em que $q \neq p$. Representa a distância entre o objeto x_j e o cluster vizinho mais próximo

Índices de validação: Relativo - Silhueta

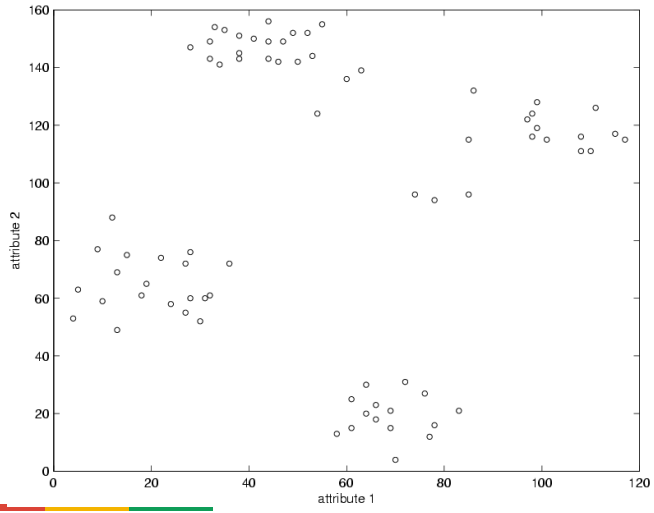


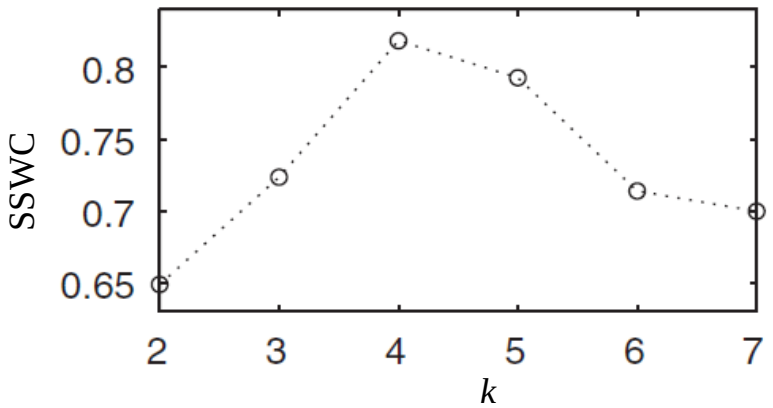
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(0).

Índices de validação: Relativo - Silhueta

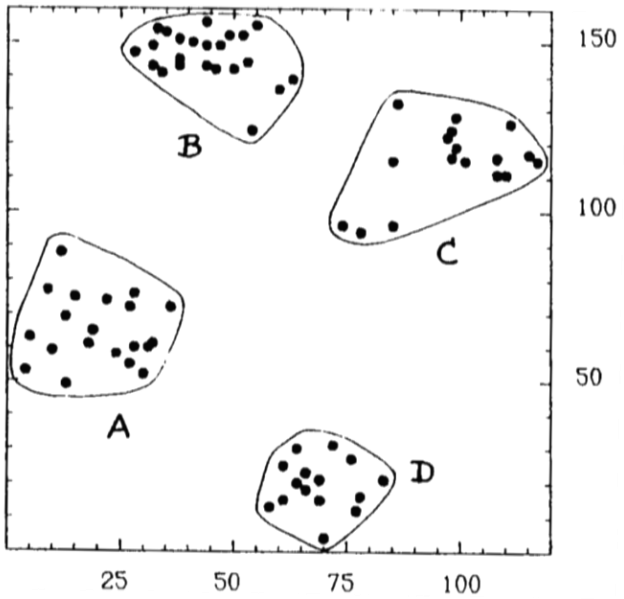
Relembrando a subjetividade do problema:

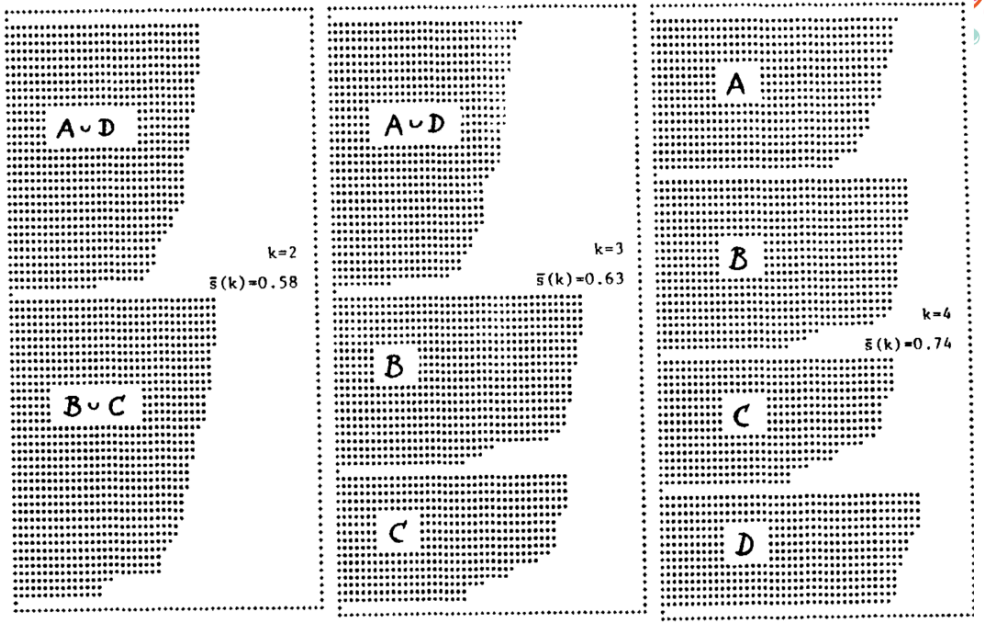
- Quantos grupos abaixo?
- Sob a perspectiva deste **critério** (SSWC) temos: *k**=4





Índices de validação: Relativo - Silhueta





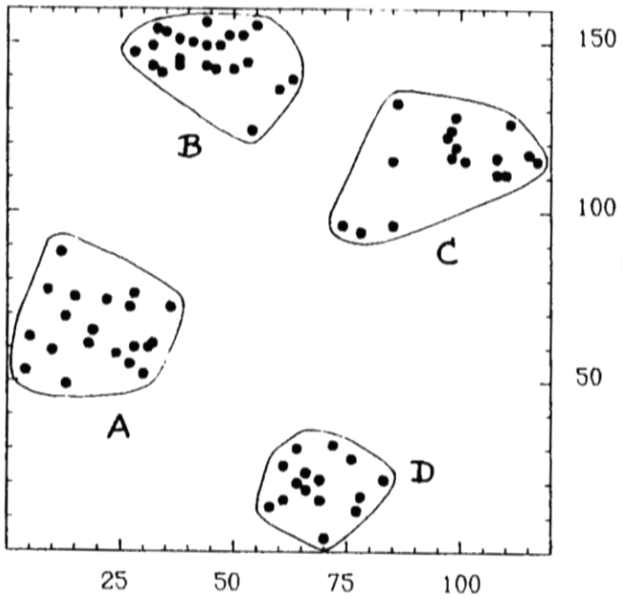
$k = 2$
$\bar{s}(k) = 0.58$

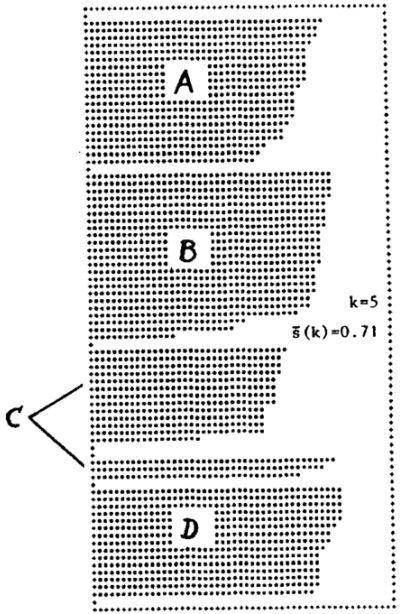
$k = 3$
$\bar{s}(k) = 0.63$

$k = 4$
$\bar{s}(k) = 0.74$

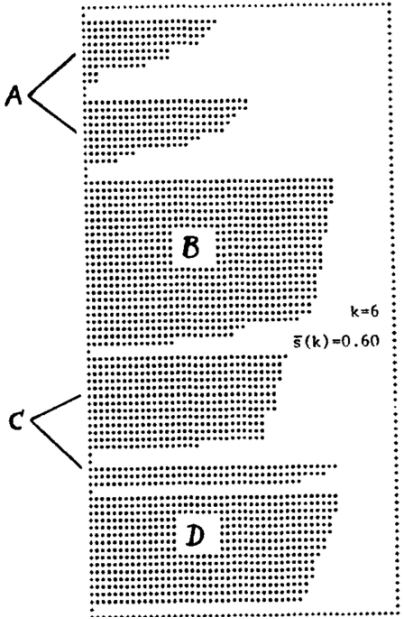
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(0).

Índices de validação: Relativo - Silhueta





$k = 5$
$\bar{s}(k) = 0.71$



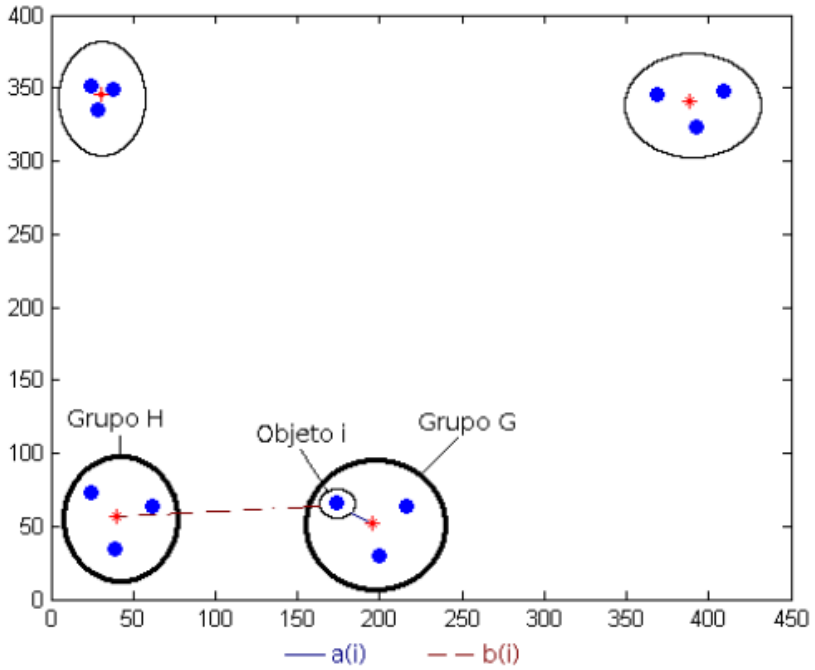
$k = 6$
$\bar{s}(k) = 0.60$

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(0).

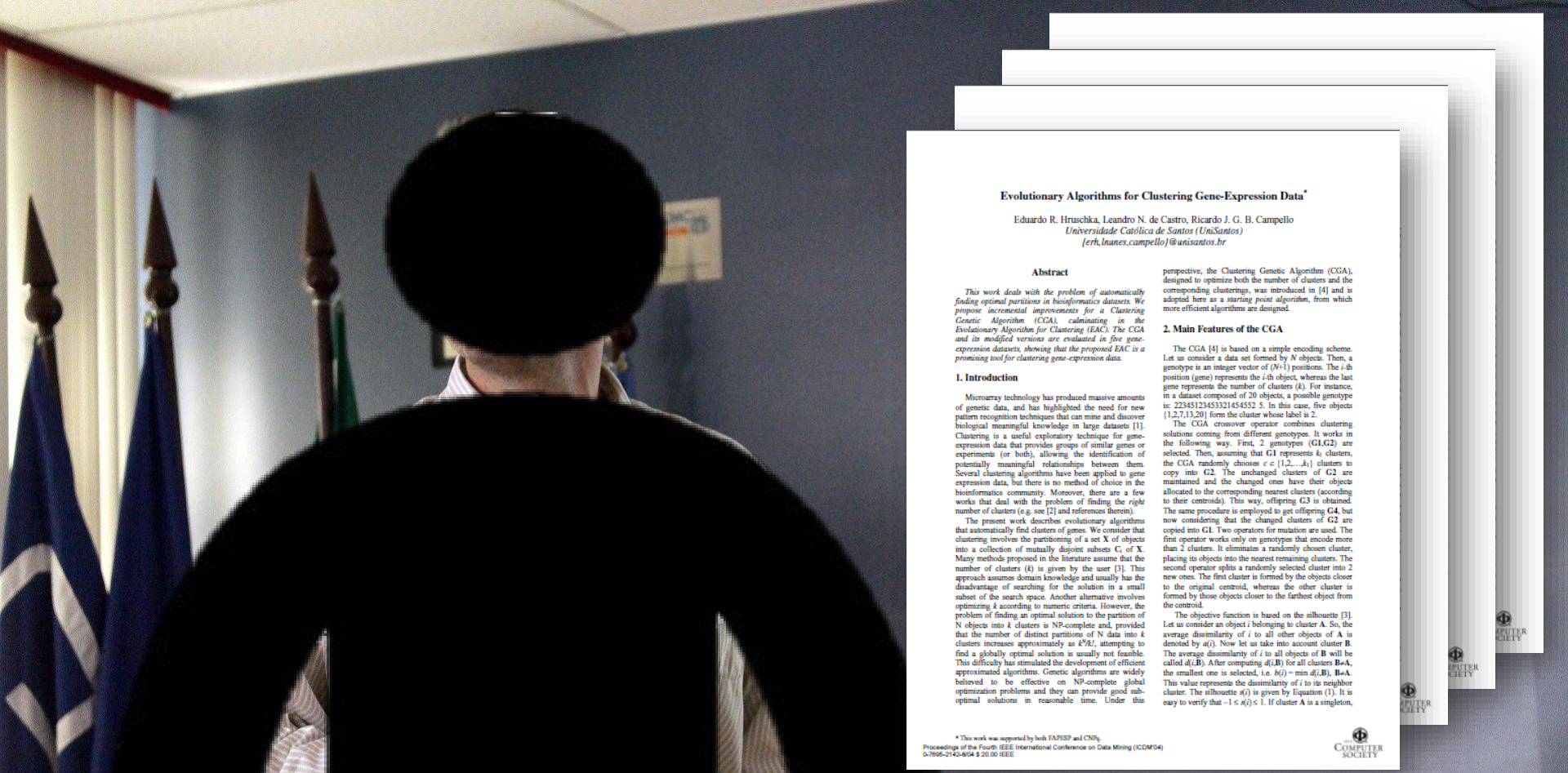
Índices de validação: Relativo – Silhueta Simplificada

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$SWC = \frac{1}{N} \sum_{i=1}^{N} s(i)$$



Silhueta Simplificada: $a(i)$ e $b(i)$ são calculados como a distância do i-ésimo objeto ao centróide do cluster em questão - O(N).



Evolutionary Algorithms for Clustering Gene-Expression Data*

Eduardo R. Hruschka, Leandro N. de Castro, Ricardo J. G. B. Campello
Universidade Católica de Santos (UniSantos)
{erh,lnunes,campello}@unisantos.br

Abstract

This work deals with the problem of automatically finding optimal partitions in bioinformatics datasets. We propose incremental improvements for a Clustering Genetic Algorithm (CGA), culminating in the Evolutionary Algorithm for Clustering (EAC). The CGA and its modified versions are evaluated in five gene-expression datasets, showing that the proposed EAC is a promising tool for clustering gene-expression data.

1. Introduction

Microarray technology has produced massive amounts of genetic data, and has highlighted the need for new pattern recognition techniques that can mine and discover biological meaningful knowledge in large datasets [1]. Clustering is a useful exploratory technique for gene-expression data that provides groups of similar genes or experiments (or both), allowing the identification of potentially meaningful relationships between them. Several clustering algorithms have been applied to gene expression data, but there is no method of choice in the bioinformatics community. Moreover, there are a few works that deal with the problem of finding the *right* number of clusters (e.g. see [2] and references therein).

The present work describes evolutionary algorithms that automatically find clusters of genes. We consider that clustering involves the partitioning of a set **X** of objects into a collection of mutually disjoint subsets C_i of **X**. Many methods proposed in the literature assume that the number of clusters (k) is given by the user [3]. This approach assumes domain knowledge and usually has the disadvantage of searching for the solution in a small subset of the search space. Another alternative involves optimizing k according to numeric criteria. However, the problem of finding an optimal solution to the partition of N objects into k clusters is NP-complete and, provided that the number of distinct partitions of N data into k clusters increases approximately as $k^N/k!$, attempting to find a globally optimal solution is usually not feasible. This difficulty has stimulated the development of efficient approximated algorithms. Genetic algorithms are widely believed to be effective on NP-complete global optimization problems and they can provide good sub-optimal solutions in reasonable time. Under this perspective, the Clustering Genetic Algorithm (CGA), designed to optimize both the number of clusters and the corresponding clusterings, was introduced in [4] and is adopted here as a *starting point algorithm*, from which more efficient algorithms are designed.

2. Main Features of the CGA

The CGA [4] is based on a simple encoding scheme. Let us consider a data set formed by N objects. Then, a genotype is an integer vector of $(N+1)$ positions. The i-th position (gene) represents the i-th object, whereas the last gene represents the number of clusters (k). For instance, in a dataset composed of 20 objects, a possible genotype is: 22345123453321454552 5. In this case, five objects {1,2,7,13,20} form the cluster whose label is 2.

The CGA crossover operator combines clustering solutions coming from different genotypes. It works in the following way. First, 2 genotypes (**G1**,**G2**) are selected. Then, assuming that **G1** represents k_1 clusters, the CGA randomly chooses $c \in \{1,2,\ldots,k_1\}$ clusters to copy into **G2**. The unchanged clusters of **G2** are maintained and the changed ones have their objects allocated to the corresponding nearest clusters (according to their centroids). This way, offspring **G3** is obtained. The same procedure is employed to get offspring **G4**, but now considering that the changed clusters of **G2** are copied into **G1**. Two operators for mutation are used. The first operator works only on genotypes that encode more than 2 clusters. It eliminates a randomly chosen cluster, placing its objects into the nearest remaining clusters. The second operator splits a randomly selected cluster into 2 new ones. The first cluster is formed by the objects closer to the original centroid, whereas the other cluster is formed by those objects closer to the farthest object from the centroid.

The objective function is based on the silhouette [3]. Let us consider an object i belonging to cluster **A**. So, the average dissimilarity of i to all other objects of **A** is denoted by $a(i)$. Now let us take into account cluster **B**. The average dissimilarity of i to all objects of **B** will be called $d(i,\mathbf{B})$. After computing $d(i,\mathbf{B})$ for all clusters **B**≠**A**, the smallest one is selected, i.e. $b(i) = \min d(i,\mathbf{B})$, **B**≠**A**. This value represents the dissimilarity of i to its neighbor cluster. The silhouette $s(i)$ is given by Equation (1). It is easy to verify that $-1 \le s(i) \le 1$. If cluster **A** is a singleton,

* This work was supported by both FAPESP and CNPq.

Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)
0-7695-2142-8/04 $ 20.00 IEEE

E. R. Hruschka, L. N. de Castro and R. J. G. B. Campello, "Evolutionary algorithms for clustering gene-expression data," *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, 2004, pp. 403-406. doi: 10.1109/ICDM.2004.10073
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1410321&isnumber=30565

Índices de validação: Externos

A maneira quantitativa para validação é alcançada através de algum tipo de **índice**. Há 3 tipos de **índices/critérios de validade**:

- **Internos**: Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados.
- **Relativos**: Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa.
- **Externos**: Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação a priori na forma de uma solução de agrupamento esperada ou conhecida.

Índices de validação: Externos

Estudaremos os índices mais usados (Rand e Jaccard). Adotaremos a seguinte terminologia:

- grupos da **partição de referência** (*golden truth*) → "**classes**"
- grupos da **partição sob avaliação** → **clusters (grupos)**

Podemos então definir as grandezas de interesse:

a: No. de pares da mesma classe e do mesmo cluster

b: No. de pares da mesma classe e de clusters distintos

c: No. de pares de classes distintas e do mesmo cluster

d: No. de pares de classes e clusters distintos

Índices de validação: Externos – Rand Index

$$RI = \frac{a+d}{a+b+c+d}$$

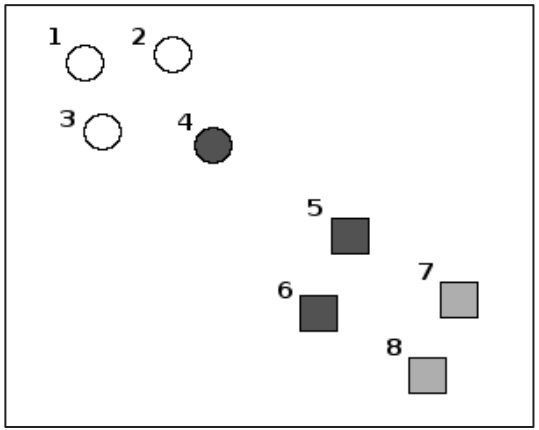
Número de pares de objetos:

a: da mesma classe e do mesmo cluster (grupo)

b: da mesma classe e de clusters distintos

c: de classes distintas e do mesmo cluster

d: de classes distintas e de clusters distintos



2 Classes (Círculos e Quadrados)
3 Clusters (Preto, Branco e Cinza)

a = 5; b = 7; c = 2; d = 14

RI = 5+14/(5+7+2+14) = 0.6785

Índices de validação: Externos – Rand Index

favorece a comparação de partições com níveis mais elevados de granularidade, i.e., apresenta valores mais elevados ao comparar partições com mais grupos.

Por quê?

- mesmo peso para objetos agregados (termo **a**) ou separados (**d**);
- termo **d** tende a dominar o índice;
- quanto mais grupos, mais pares pertencem a grupos distintos;
  - isso é válido em qualquer uma das duas partições;
  - probabilidade / incidência de pares em comum é maior.

Índices de validação: Externos – Jaccard

Elimina o termo **d** sob a ótica de que um agrupamento é uma coleção de agregações de pares de objetos (separações sendo apenas uma consequência):

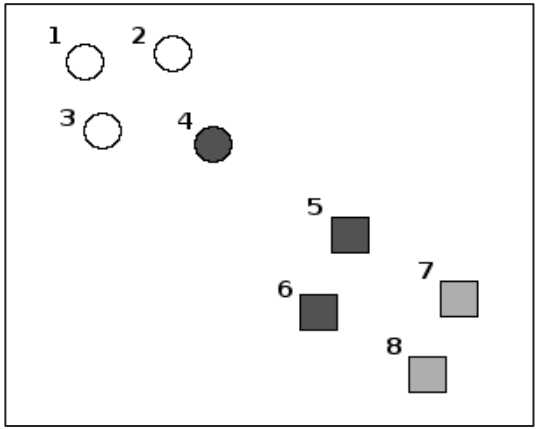
$$Jc = \frac{a}{a+b+c}$$

Número de pares de objetos:

a: da mesma classe e do mesmo cluster

b: da mesma classe e de clusters distintos

c: de classes distintas e do mesmo cluster



2 Classes (Círculos e Quadrados)
3 Clusters (Preto, Branco e Cinza)

a = 5; b = 7; c = 2

Jc = 5/(5+7+2) = 0.3571

Existem vários outros critérios

-	Criterion	Complexity
	Calinski-Harabasz (VRC)	$O(nN)$
	Davies-Bouldin (DB)	$O(n(k^2+N))$
	Dunn	$O(nN^2)$
	Silhouette Width Criterion (SWC)	$O(nN^2)$
	Alternative Silhouette (ASWC)	$O(nN^2)$
	Simplified Silhouette (SSWC)	$O(nNk)$
	Alternative Simplified Silhouette (ASSWC)	$O(nNk)$
	PBM	$O(n(k^2+N))$
	C-Index	$O(N^2(n+log_2N))$
	Gamma	$O(nN^2+N^4/k])$
	G(+)	$O(nN^2+N^4/k])$
	Tau	$O(nN^2+N^4/k])$
	Point-Biserial	$O(nN^2)$
	C/$\sqrt{k}$	$O(nN)$
*	Trace(W)	$O(nN)$
*	Trace(CovW)	$O(nN)$
*	Trace($W^{-1}B$)	$O(n^2N+n^3)$
*	|T|/|W|	$O(n^2N+n^3)$
*	Nlog(|T|/|W|)	$O(n^2N+n^3)$
*	k^2W	$O(n^2N+n^3)$
*	log(SSB/SSW)	$O(n(k^2+N))$
*	Ball-Hall	$O(nN)$
*	McClain-Rao	$O(nN^2)$

Vendramin, Campello, Hruschka "*Relative Clustering Validity Criteria: A Comparative Overview*" **Statistical Analysis and Data Mining**, Vol. 3, p. 209-235, 2010.

Referências

1. Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
2. Everitt, B. S., Landau, S., and Leese, M., Cluster Analysis, Arnold, 4th Edition, 2001.
3. Tan, P.-N., Steinbach, M., and Kumar, V., Introduction to Data Mining, Addison-Wesley, 2006
4. Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 2005.
5. Wu, X. and Kumar, V., The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC, 2009
6. D. Steinley, K-Means Clustering: A Half-Century Synthesis, British J. of Mathematical and Stat. Psychology, V. 59, 2006

Questions and Feedback

Thank you!

Obrigado !

Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá
@machine learning Brasil

@vfcarida