

## *Machine Learning*



## *Clustering, Support Vector Machines* e processamento em linguagem natural

Bloco 1

Lucas Claudino



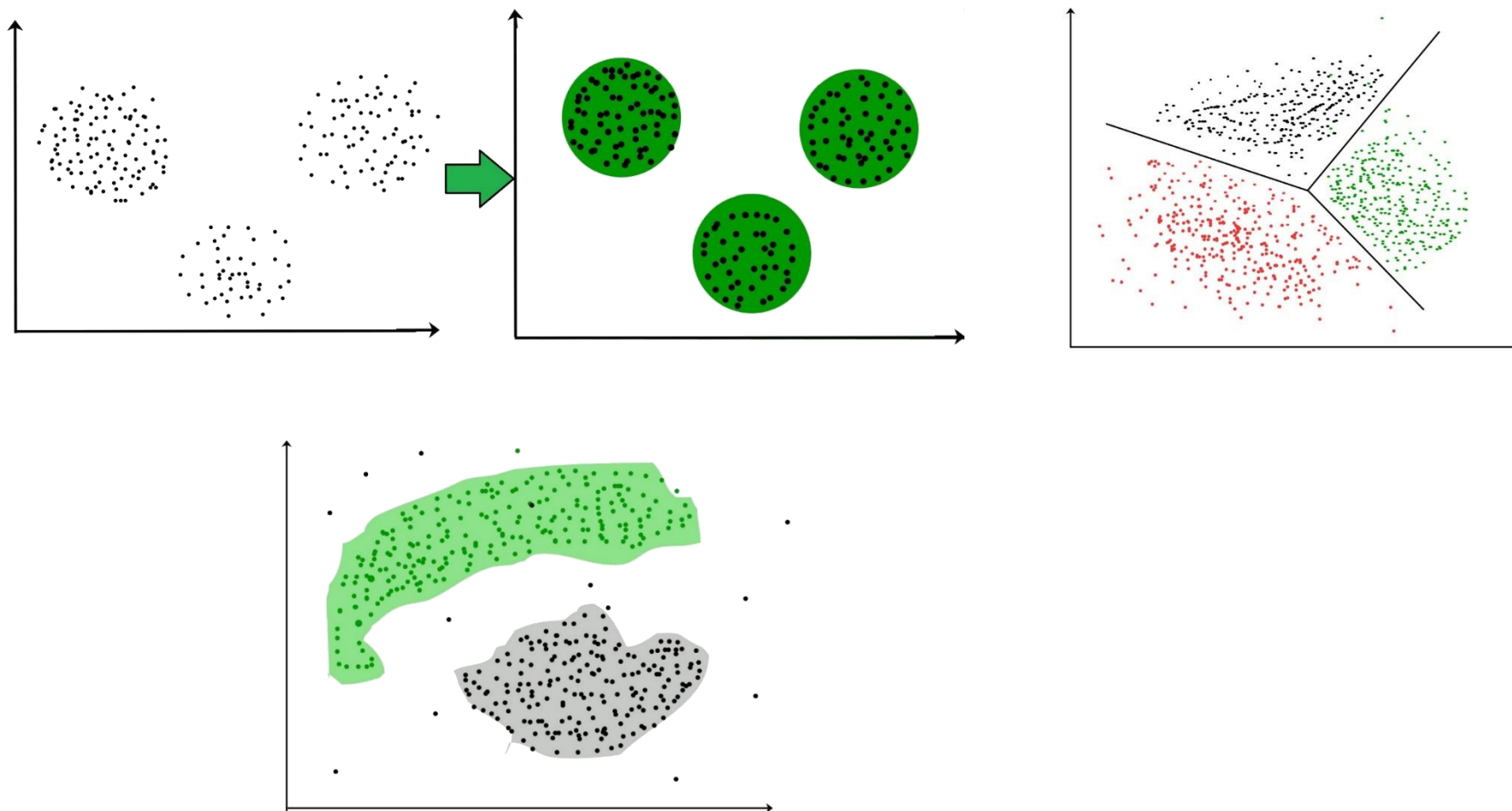


## ► *Clustering*

- **Objetivo:** encontrar grupos com características similares.
- Critérios de separação:
  - Compactação.
  - Encadeamento ou ligação.
  - Separação espacial.
- Algoritmos de agrupamento:
  - Hierárquico (aglomerativo ou divisivo).
  - Baseados em erro quadrático (k-means).
  - Baseados em densidade.
  - Baseados em grafo.
  - Baseados em redes neurais.

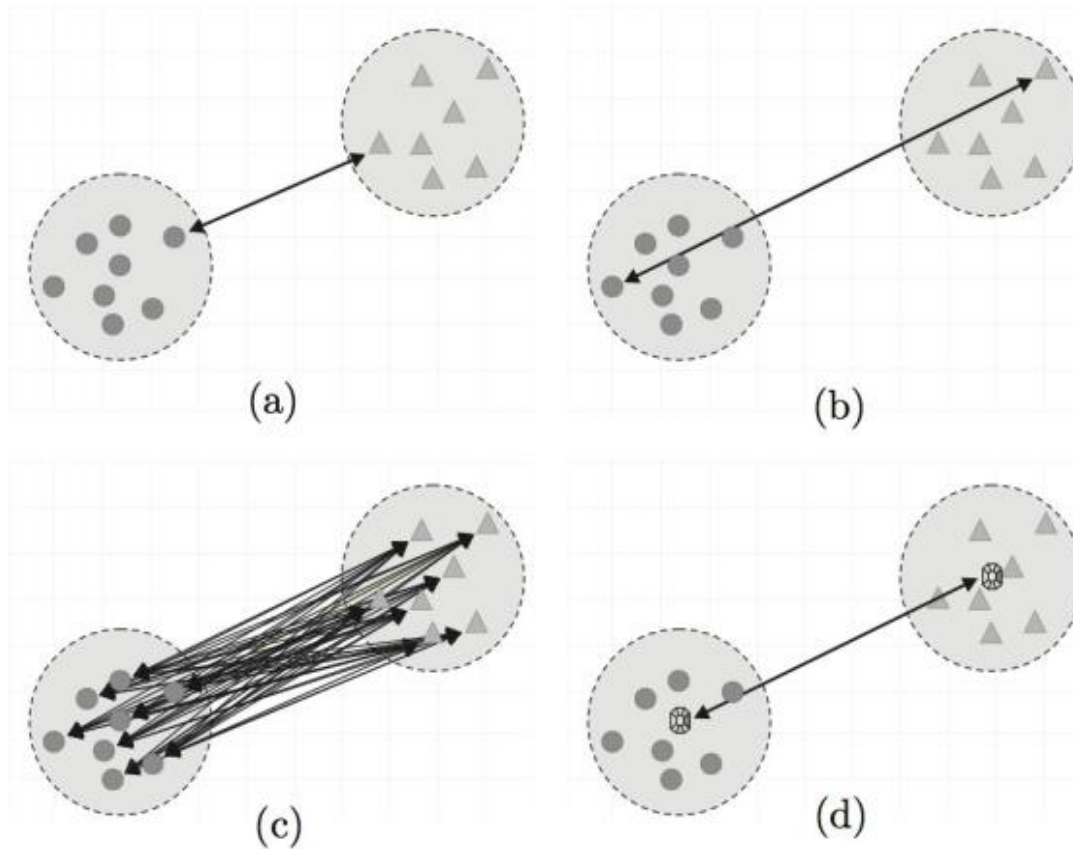
## ► Clustering

Figura 1 – Exemplos de diferentes configurações espaciais de grupos



## ► *Clustering* – critérios para distância

Figura 2 – Distâncias a) mínima; b) máxima; c) média; d) entre centroides

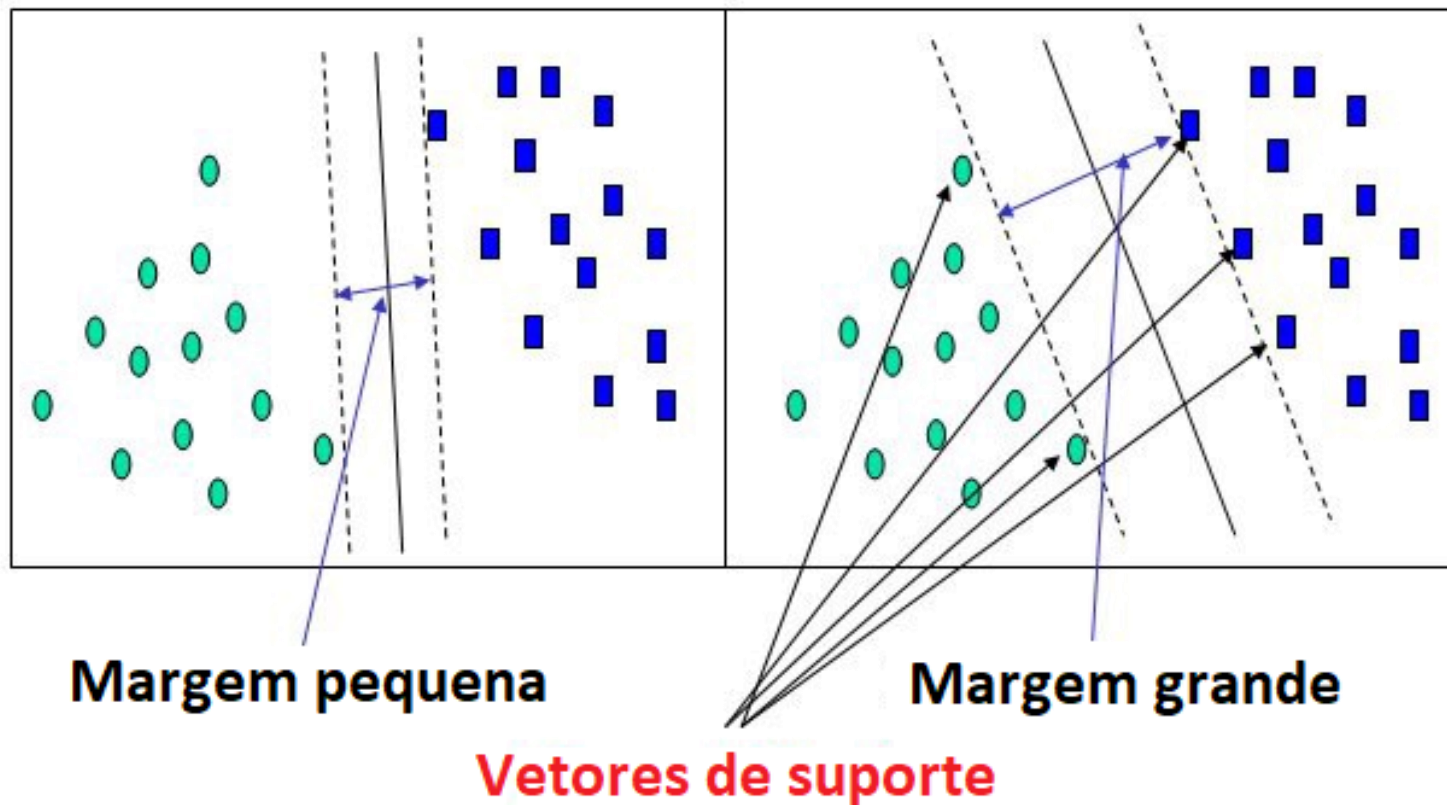


Fonte: adaptada de Faceli (2011, p. 211).



## ► *Support Vector Machines*

Figura 3 – Estrutura de otimização de uma máquina de vetor de suporte



## ► *Support Vector Machines*

- **Objetivo:** encontrar probabilidade para uma sequência de palavras.
- **N-grama:** algoritmo mais simples.

$$P(\textit{palavra}(s) | \textit{frase})$$

- Aplicações: interação entre homem e máquina.
  - Reconhecimento de voz.
  - Identificação de idioma.

## ***Clustering: hierárquico, $k$ -means e densidade***

**Bloco 2**

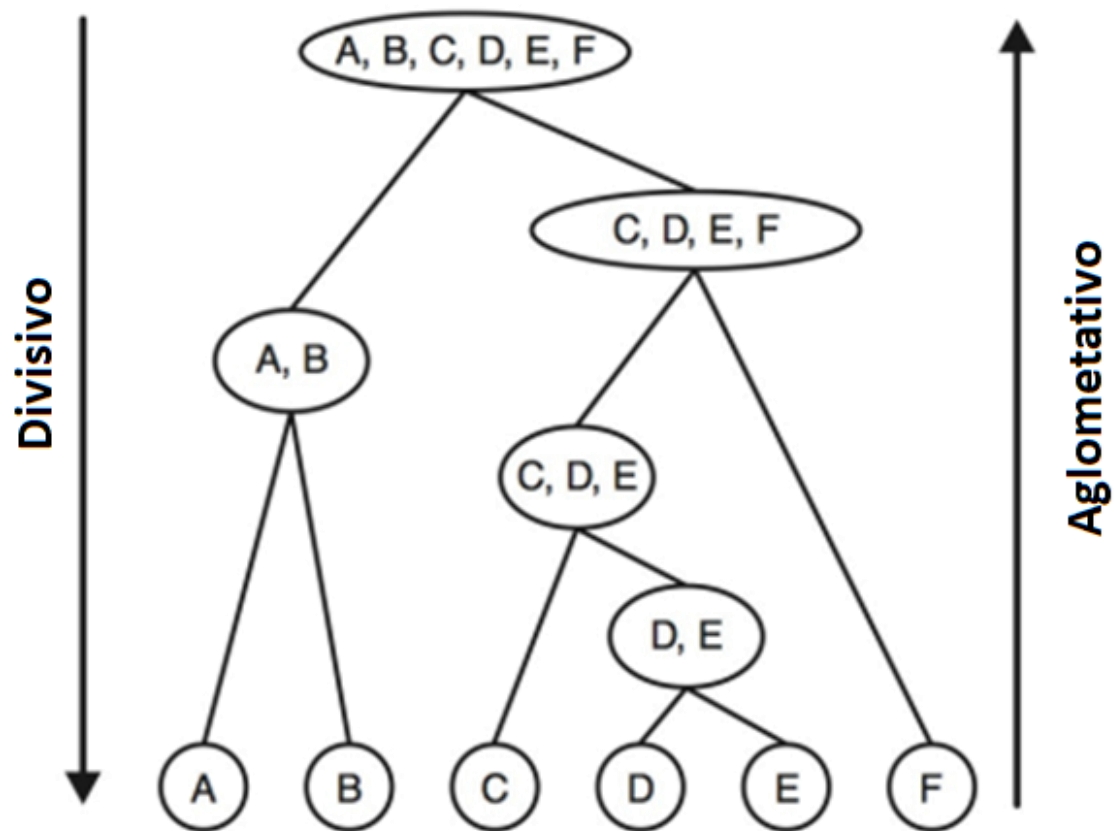
Lucas Claudino





## ► Agrupamento hierárquico

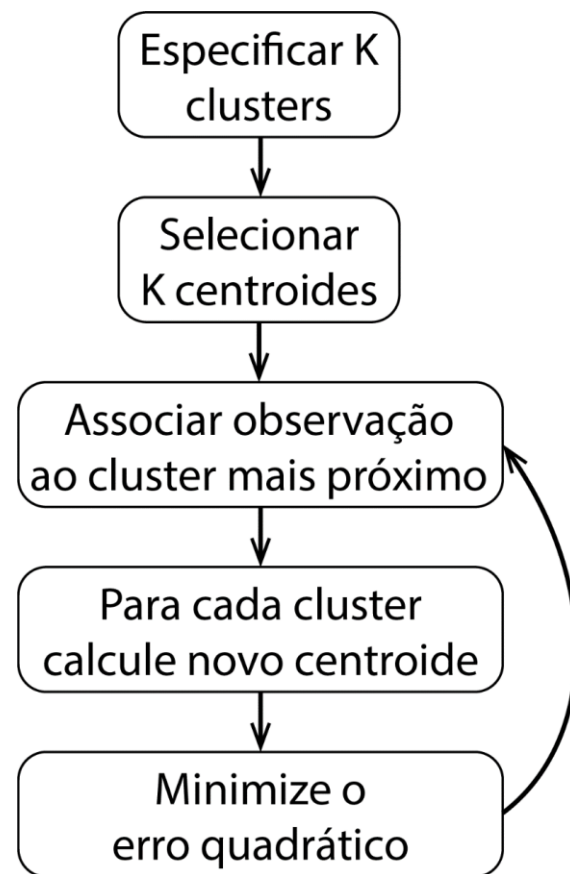
Figura 4 – Agrupamento do tipo aglomerativo



Fonte: Faceli (2011, p. 210).

## ► *K-means*

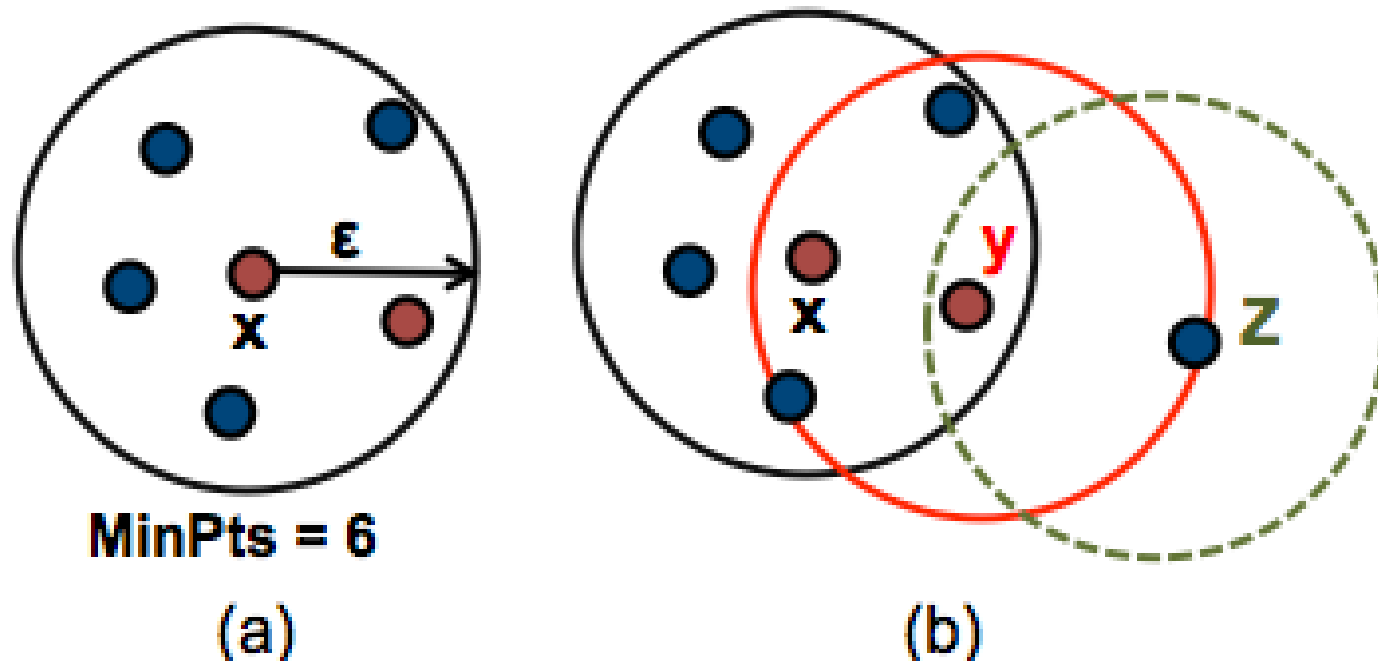
**Figura 5 – Sequência lógica de implementação do algoritmo k-means**



Fonte: elaborada pelo autor.

## ► Baseados em densidade

Figura 6 - Disposição de pontos e classificação via DBSCAN



Fonte: Ester *et al.* (1996).

# PÓS-GRADUAÇÃO

## Teoria em prática

### Bloco 3

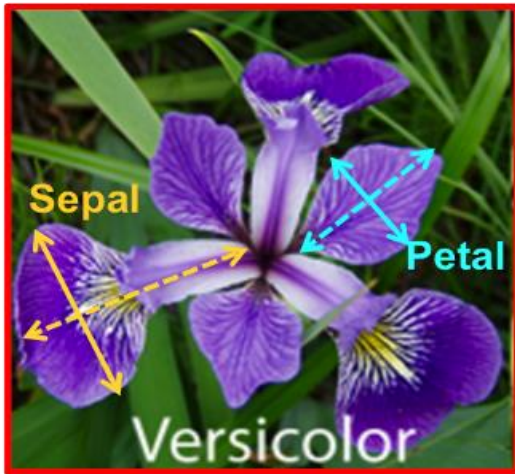
Lucas Claudino



## ► Teoria em prática: aplicação do *k-means*

- Problema Iris.

Figura 7 – Plantas setosa, versicolor e virgínica



Fonte: <https://mc.ai/visualization-and-understanding-iris-dataset/>. Acesso em: 14 ago. 2019.



## ► Teoria em prática

Figura 7 – Exemplo de implementação de algoritmo k-means para resolução do problema Iris.

```
kMeans1 <- function(dados, k=2){  
  # função que calcula a distância euclidiana  
  euc.dist <- function(x1, x2) sum((x1 - x2) ^ 2)  
  rotulo = 1:k      # labels  
  rownames(dados)[nrow(dados)] = 1  
  for(i in 1:nrow(dados)){  
    rownames(dados)[i] <- sample(rotulo, 1)  
  } # centroides aleatorios  
  centroids <- colMeans(dados[rownames(dados) == 1, ])  
  for(j in 2:k){  
    centroids <- rbind(centroids,  
      |         |         | colMeans(dados[rownames(dados) == j, ]))  
  }  
  # identifica o centroide de cada grupo  
  rownames(centroids) = 1:k
```

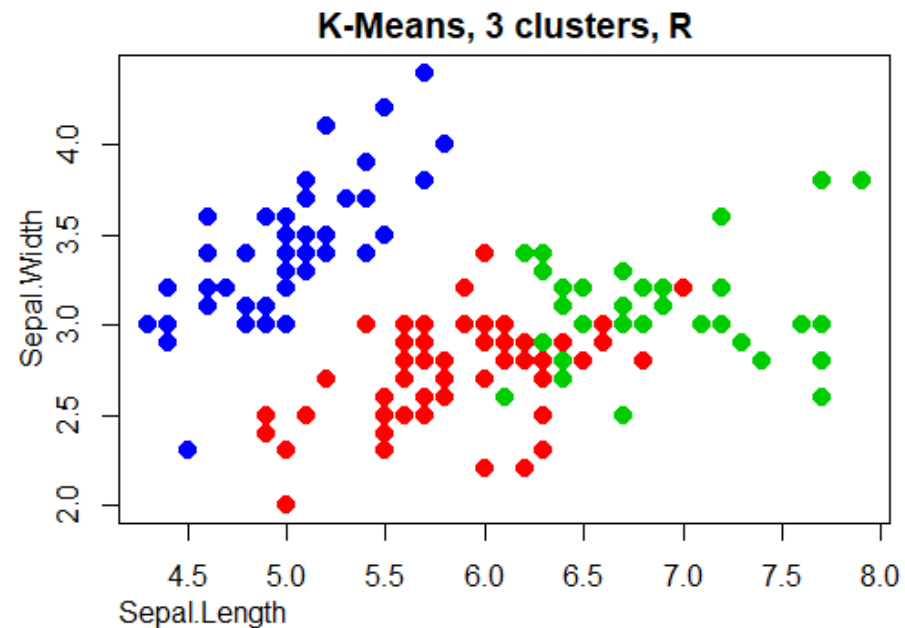
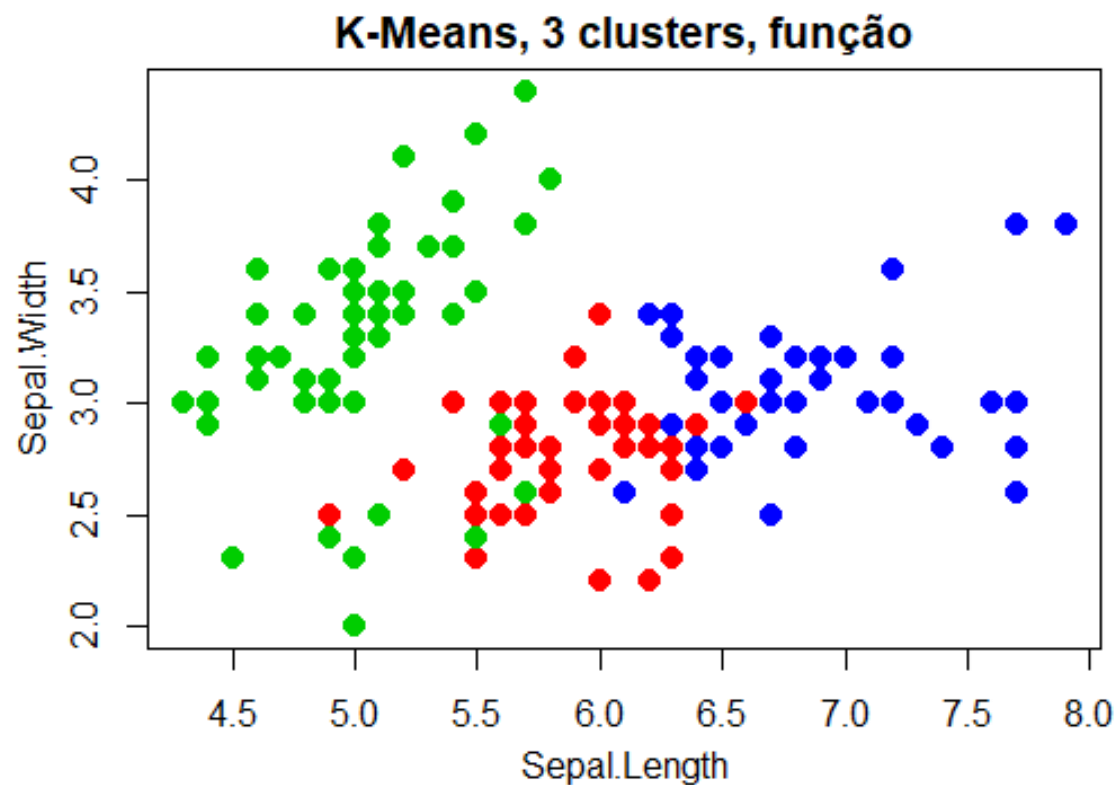
## ► Teoria em prática

Figura 8 – Exemplo de implementação de algoritmo k-means para resolução do problema Iris (2)

```
for(i in 1:nrow(dados)){
  distancias = NULL
  for(j in 1:k){
    distancias[j] = euc.dist(dados[i,], centroids[j,])
  }
  names(distancias) = 1:k
  rownames(dados)[i] = as.numeric(names(distancias[distancias == min(distancias)]))
  # recalcula as medias
  centroids <- colMeans(dados[rownames(dados) == 1, ])
  for(z in 2:k){
    centroids <- rbind(centroids, colMeans(dados[
      rownames(dados) == z, ]))
  }
  # centroides
  return(list(centroides = centroids, grupo1 = dados[
    rownames(dados) == 1, ],
    grupo2 = dados[rownames(dados) == 2, ],
    grupo3 = dados[rownames(dados) == 3, ],
    clusters = as.numeric(rownames(dados)))))
```

## ► Teoria em prática

Figura 9 – Resultado da classificação utilizando algoritmo k-means



Fonte: elaborada pelo autor.

# PÓS-GRADUAÇÃO

## Dica do professor

**Bloco 4**

Lucas Claudino





## ► Dica do professor

- Ferramenta para visualização de N-gramas.
- Probabilidade de ocorrência de N-gramas nos livros do Google.
- <https://books.google.com/ngrams/>






## ► Referências

ESTER, Martin *et al.* A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings Of The Second International Conference On Knowledge Discovery And Data Mining. **Oregon**, p. 226-231, ago./1996.

FACELI, K. *et al.* **Inteligência artificial**: uma abordagem de aprendizado de máquina. São Paulo: LTC Editora, 2011.



## ► Referências

GANDHI, Rohith. Support Vector Machine: Introduction to Machine Learning Algorithms. **Towards Data Science**, 7 de junho de 2018. Disponível em: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Acesso em: 2 set. 2019.

PRIY, Surya. Clustering in Machine Learning. **GeeksforGeeks**, 2018. Disponível em: <https://www.geeksforgeeks.org/clustering-in-machine-learning/>. Acesso em: 2 set. 2019.

SHARMA, Pranjal. Visualization and understanding: Iris Dataset. **Mc.ai**, 16 de junho de 2019. Disponível em: <https://mc.ai/visualization-and-understanding-iris-dataset/>. Acesso em: 14 ago. 2019.

