



Aprendizagem não supervisionada





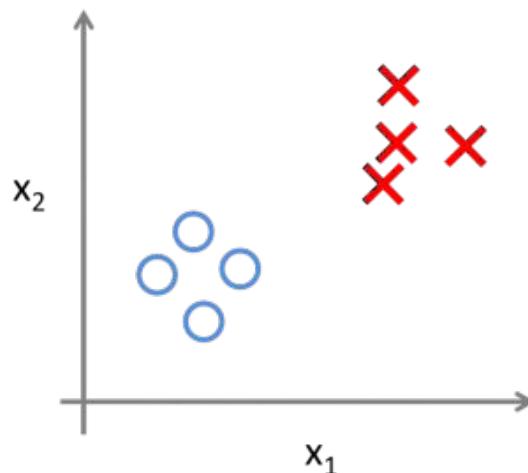
“Você pode ter os dados sem informações, **mas** você não pode ter informações sem dados”.



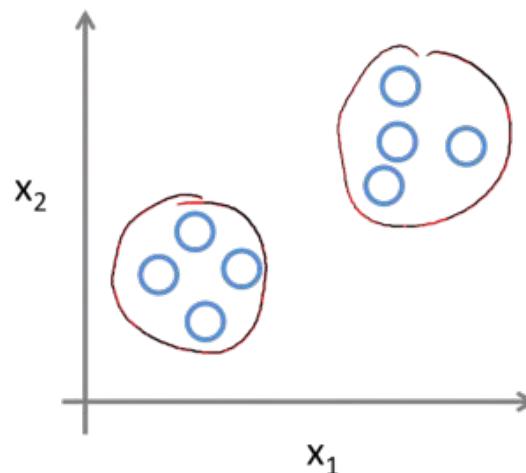
Aprendizagem não supervisionada

Introdução

Supervised Learning



Unsupervised Learning



Introdução

- ◊ Classificar novos dados não rotulados
- ◊ Formação de grupos
- ◊ Exemplos:
 - Agrupamento de clientes
 - Cocktail party

Introdução

- ◊ Podemos ganhar alguma percepção da natureza (ou estrutura) dos dados.
- ◊ Podemos Identificar padrões implícitos
- ◊ Procurar por outliers e detectar anomalias

Introdução

- ◊ Objetivo: Amostras dentro de um mesmo cluster sejam muito parecidos, e amostras em clusters diferentes sejam distintos entre si.
- ◊ São utilizados medidas de similaridade entre as amostras
- ◊ Exemplos de análise de cluster são:
 - Baseado no centroíde
 - Baseado na densidade

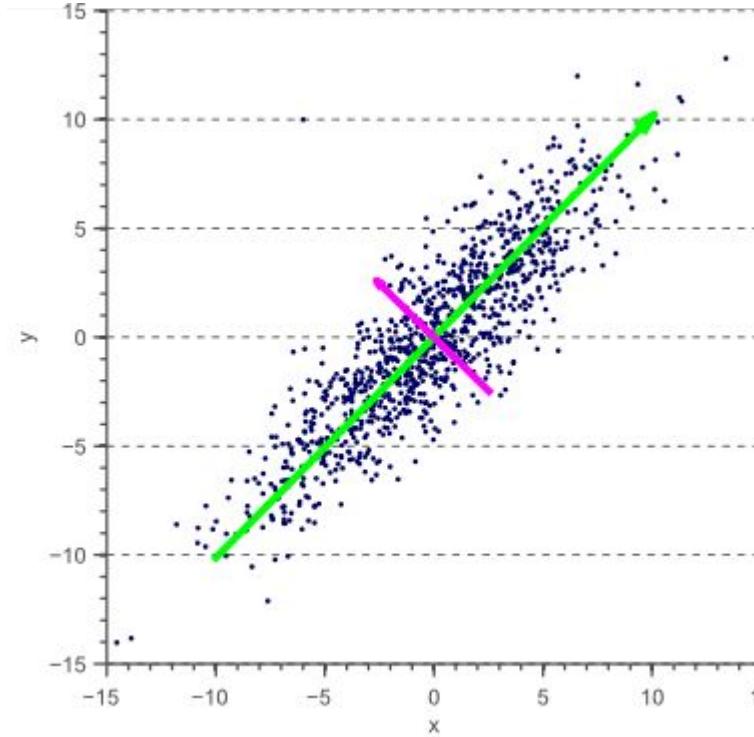


PCA



Objetivo

- ◊ Extração de atributos
- ◊ Diminuição de informação redundante
- ◊ Facilidade na computação dos modelos
- ◊ Facilidade na visualização dos dados



Exemplo de identificação de duas componentes principais



Tabela

	% Variance Explained	Cumulative % Variance Explained
PC 1	0.29	0.29
PC 2	0.18	0.47
PC 3	0.14	0.61
PC 4	0.09	0.70
PC 5	0.08	0.77
PC 6	0.05	0.82
PC 7	0.04	0.87
PC 8	0.04	0.90
PC 9	0.03	0.93
PC 10	0.02	0.96
PC 11	0.02	0.98
PC 12	0.01	0.99
PC 13	0.01	1.00



Exemplo

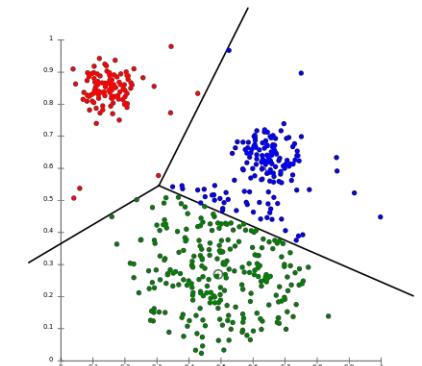
	sepal length	sepal width	petal length	petal width
0	-0.900681	1.032057	-1.341272	-1.312977
1	-1.143017	-0.124958	-1.341272	-1.312977
2	-1.385353	0.337848	-1.398138	-1.312977
3	-1.506521	0.106445	-1.284407	-1.312977
4	-1.021849	1.263460	-1.341272	-1.312977

PCA
(2 components) 

	principal component 1	princial component 2
0	-2.264542	0.505704
1	-2.086426	-0.655405
2	-2.367950	-0.318477
3	-2.304197	-0.575368
4	-2.388777	0.674767

KMEANS

Um dos mais conhecidos algoritmo de clusterização



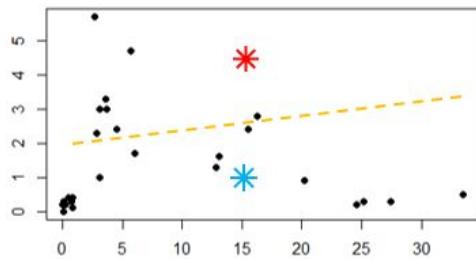


Introdução

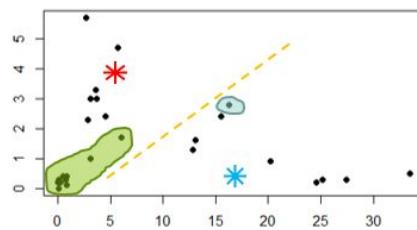
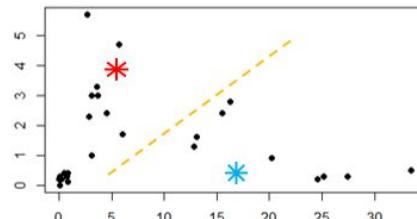
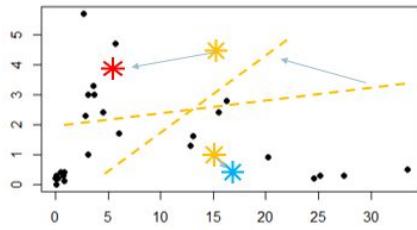
- ◊ Baseado no centroíde (representante)
- ◊ Os centróides (pontos centrais do grupo)
- ◊ **K-means** tem como objetivo agrupar os dados em C_1, C_2, \dots, C_k clusters, de acordo com a medida de distância média (**means**) dos pontos, o qual define os centróides
- ◊ As partições têm a finalidade de minimizar a distância de todos os pontos, de um cluster, aos seus respectivos centróides

Algoritmo

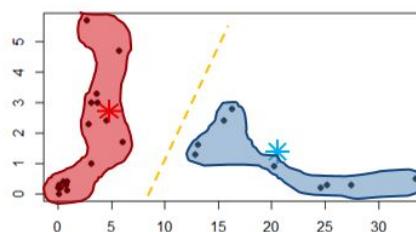
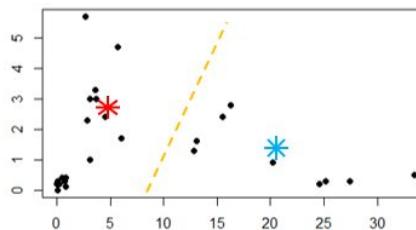
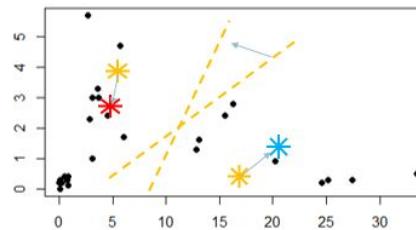
1. Escolhe randomicamente K pontos para representar os centróides iniciais
2. Agrupa todos os pontos aos seus centróides mais próximos, de acordo com a medida de distância.
3. Verifica as seguintes situações:
 - a. Se houve mudanças de grupos/novas atribuições de pontos, calcula-se novamente o vetor central de um grupo (definindo um novo centróide) e volto para a etapa 2.
 - b. Caso não haja mudanças, o algoritmo é finalizado.



1. Defino k=2 centróides iniciais e defino os clusters

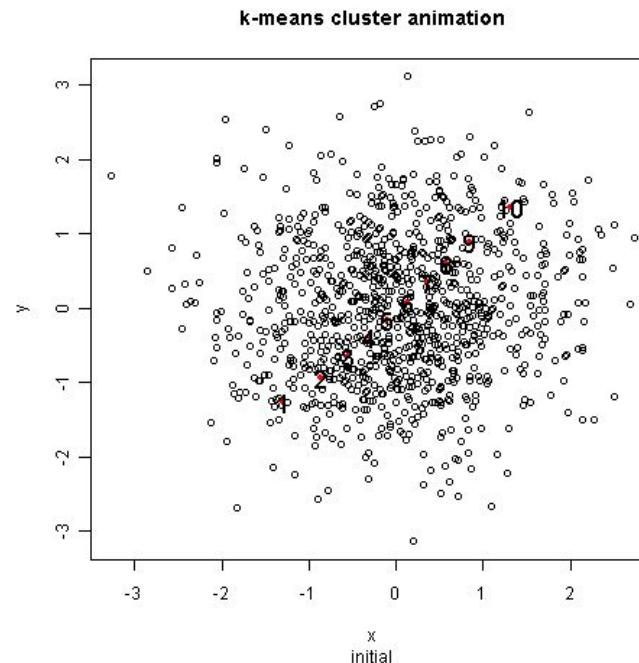


2. Recalcula o vetor central, definindo novos centróides, e assim atribui os pontos aos clusters adequados.



3. Nova disposição de centróides ocorre. Porém, nos últimos passos, não há mais movimentação de pontos, pois os centróides realmente representam os pontos centrais do grupo agora. Assim, temos o grupo “vermelho” e “azul”.

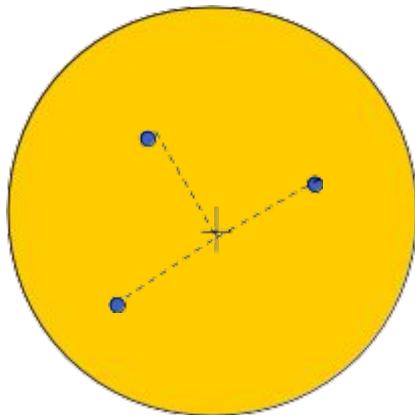
Exemplos



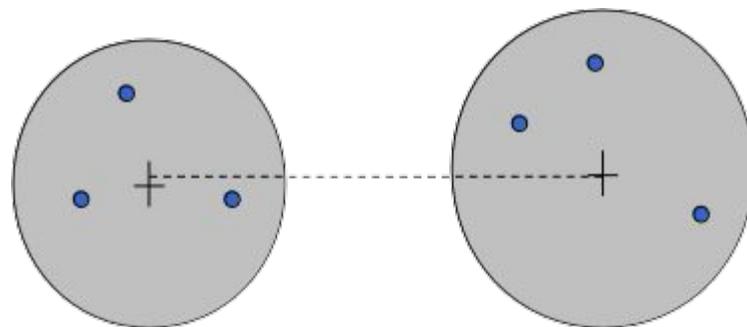


Aprendizagem não supervisionada

Como avaliar um cluster?



Coesão: mede a proximidade das amostras em relação ao centróide. É uma avaliação de um cluster (intra-cluster).



Separação: mede a qualidade de separação entre os clusters (inter-cluster).



Como escolher o K?

- ◆ Executar o algoritmo k-means para um intervalo de valores de k, calculando a Soma dos quadrados dos erros para cada clustering, obtida como:

objective function $\leftarrow J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$

number of clusters number of cases centroid for cluster j

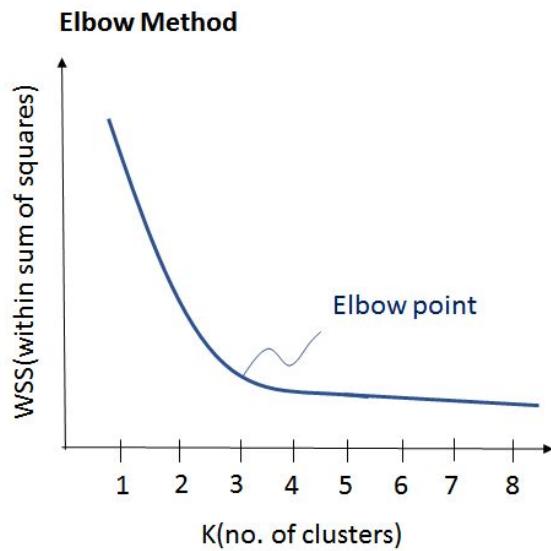
case i

Distance function

Com x um ponto de dados (vetor) e C_j o centróide do cluster K . Observe que x deve pertencer ao cluster K .



Como escolher o K?



Aplicar o método de Elbow

- ◇ Executar o algoritmo k-means para um intervalo de valores de k (e.g.: [1:8])
- ◇ O objetivo é encontrar a variação intra-cluster e minimizá-la.



Considerações

- ◊ Inicialização dos centroídes, pode provocar uma clusterização ruim
- ◊ Não há tratamento para os outliers, para isto há o **K-medoids**
- ◊ Aplique PCA se a dimensão das features for muito grande



Métricas de avaliação



Método da silhueta

- ◆ A silhueta mostra o quanto bem as amostras se posicionam dentro do cluster e quais meramente ficam em uma posição intermediária.
Assim, cada cluster é representado por uma silhueta.
- ◆ O cálculo da **Largura Média de Silhueta**, média da silhueta das amostras, é utilizado para selecionar o "melhor" número de clusters



Método da silhueta

- ◊ O Coeficiente de Silhueta é uma avaliação, em que uma pontuação mais alta de Coeficiente de Silhueta se relaciona a um modelo com clusters melhor definidos.
- ◊ O Coeficiente de Silhueta é definido para cada amostra e é composto por duas pontuações:
 - **a:** A distância média entre uma amostra e todos os outros pontos da mesma classe.
 - **b:** A distância média entre uma amostra e todos os outros pontos no próximo cluster mais próximo.
 - O Coeficiente de Silhueta s para uma única amostra é então dado como

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



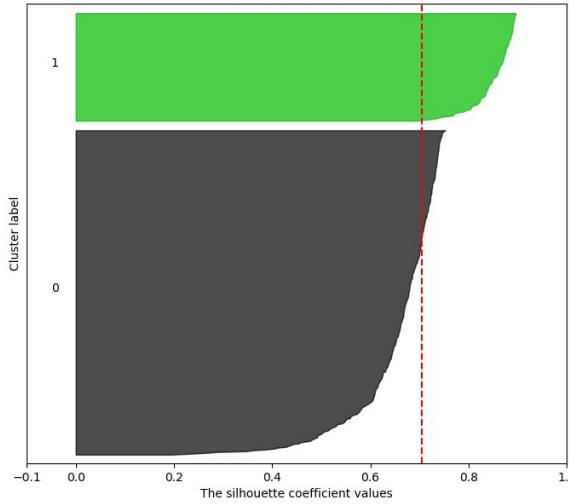
Método da silhueta

- ◊ O Coeficiente de Silhueta variam de [-1,1]:
 - **-1**: A amostra está mais próxima das amostras do cluster vizinho, mostrando que foi associada ao cluster atual erroneamente.
 - **0**: indica que a amostra está muito próxima do limite de decisão entre dois clusters vizinhos
 - **+1**: indica que a amostra está longe dos clusters vizinhos (está coeso)

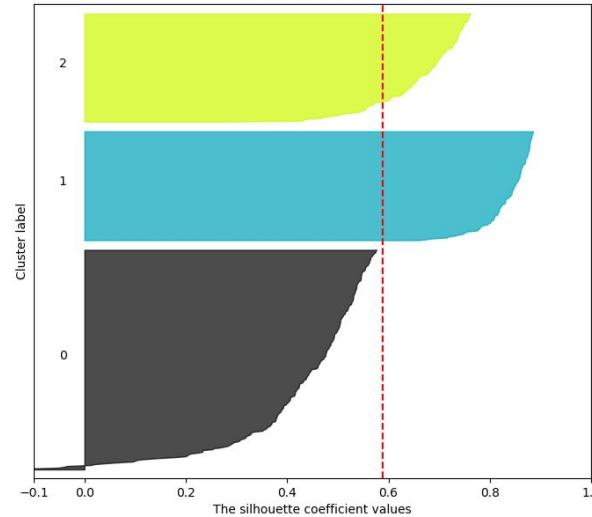


Método da silhueta

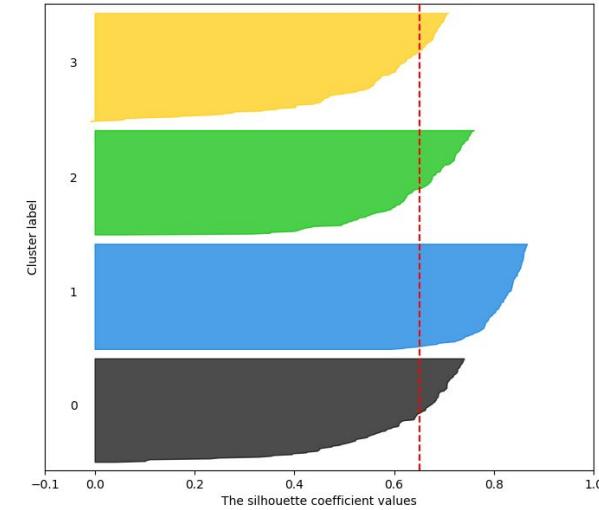
The silhouette plot for the various clusters.



The silhouette plot for the various clusters.



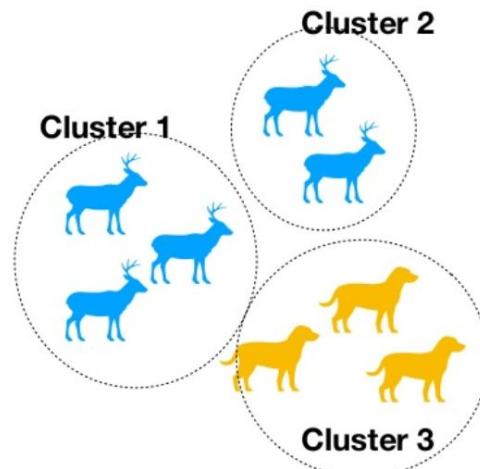
The silhouette plot for the various clusters.



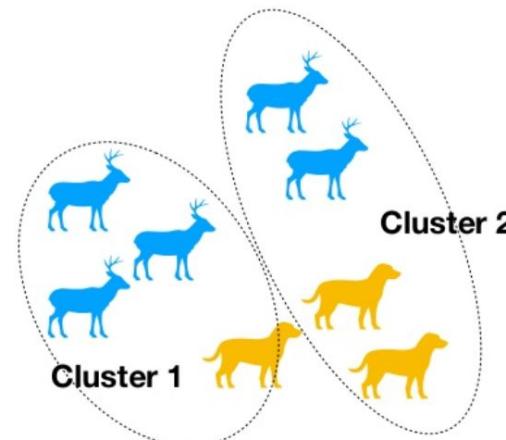


Homogeneidade

Cada cluster contém somente membros da sua classe



Good

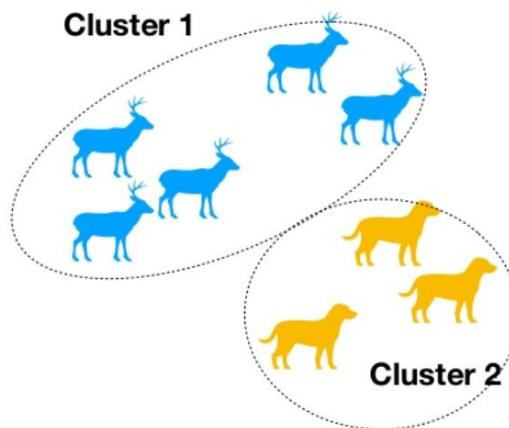


Bad

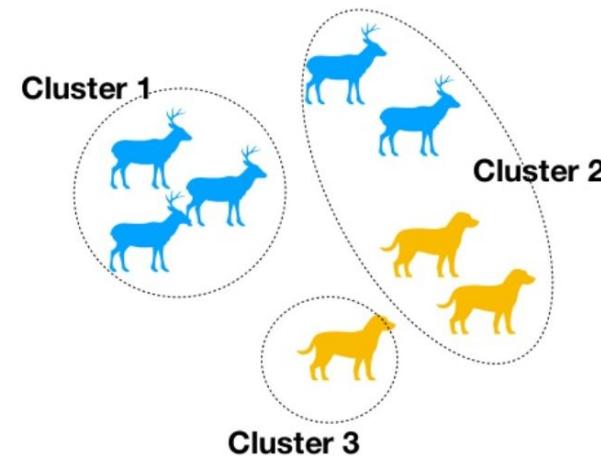


Completude

Todos os membros de uma determinada classe são atribuídos ao mesmo cluster.



Good

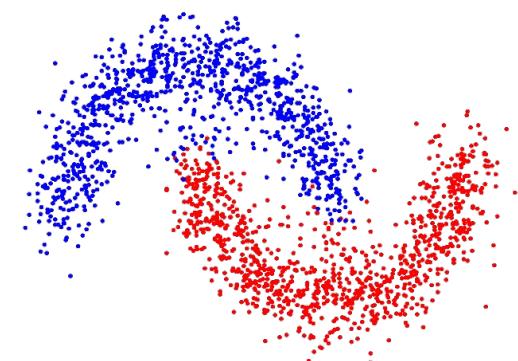


Bad



DBSCAN

Density-based spatial clustering of applications with noise





Introdução

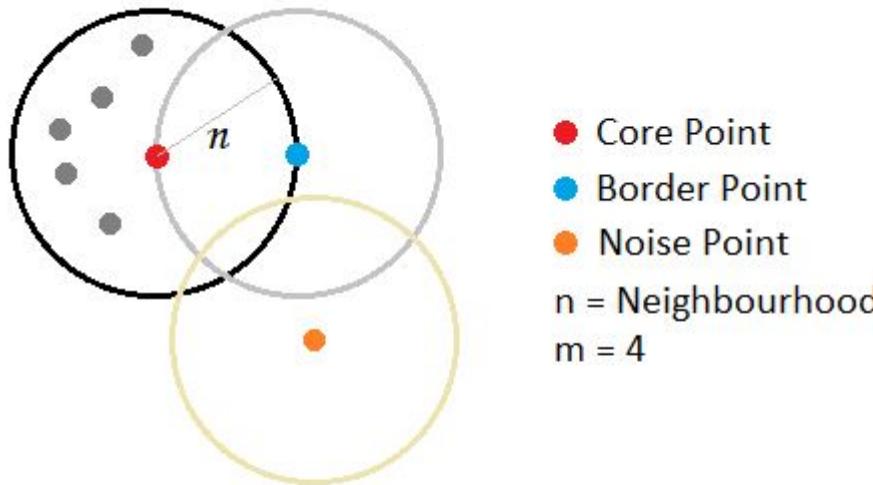
- ◊ Foi proposto em 1996
- ◊ Baseado em densidade
- ◊ Efetivo para **identificar** clusters de formato arbitrário e de diferentes tamanhos, **identificar e separar** os ruídos dos dados e **detectar clusters** “naturais” sem qualquer informação preliminar sobre os grupos.



Conceitos

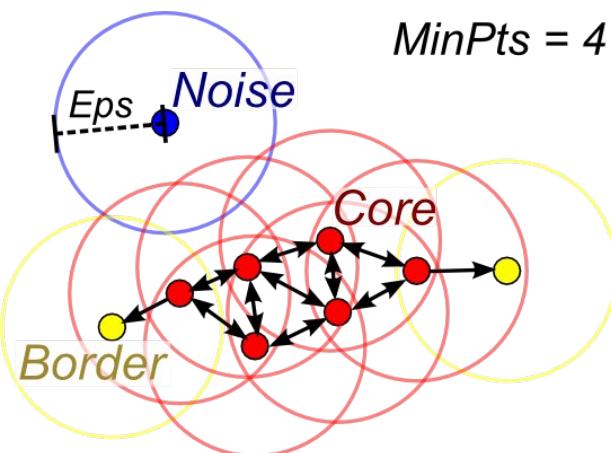
- ◊ **Vizinhança de um ponto (ϵ -vizinhança):**
 - $N(p) = \{ q \text{ em } D | \text{dist}(p,q) < \text{Raio} \}$
- ◊ **Core:**
 - Se a ϵ -vizinhança de um objeto p **contém ao menos um número mínimo, MinPts**, de objetos, então o objeto p é chamado de **ponto central** .
- ◊ **Border:**
 - Se a ϵ -vizinhança de um objeto p **contém menos que MinPts**, mas **contém algum ponto central**, então o objeto p é chamado de **ponto de borda**.
- ◊ **Noise:**
 - O ruído são o conjunto de pontos na base de dados D que não pertença a qualquer grupo Ci. Um objeto que não é ponto central nem ponto de borda, é ruído.

Conceitos



DBSCAN CLUSTERING

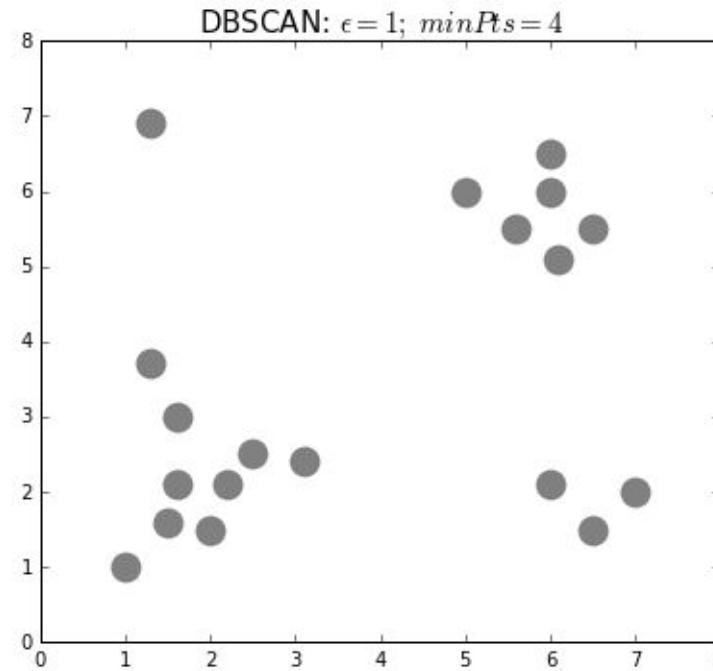
Algoritmo



- ◊ Resumindo, o agrupamento de objetos a partir de qualquer cluster de C é um processo de duas etapas.
- ◊ Na primeira, um objeto central arbitrário X do cluster 1 (X_{C1}) é identificado.
- ◊ Em seguida, todos os objetos alcançáveis por densidade a partir de X_{C1} são buscados.
- ◊ Na segunda etapa, cada cadeia de objetos partindo de X_{C1} é detectada de forma recursiva.

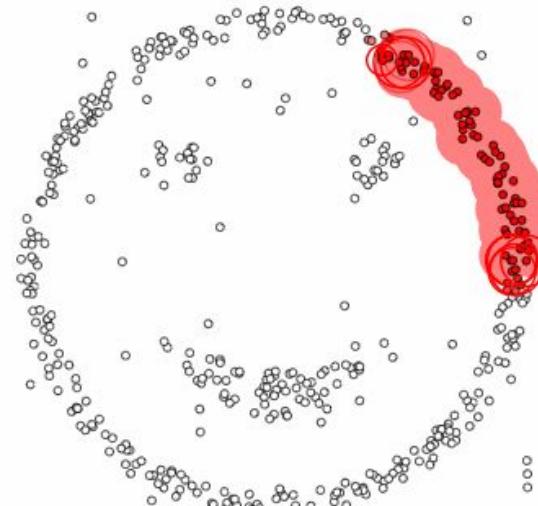


Exemplos





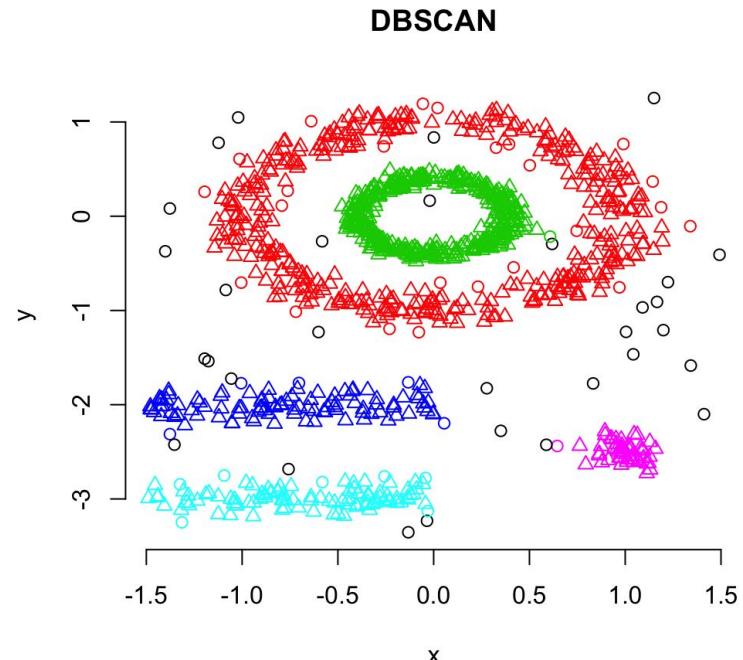
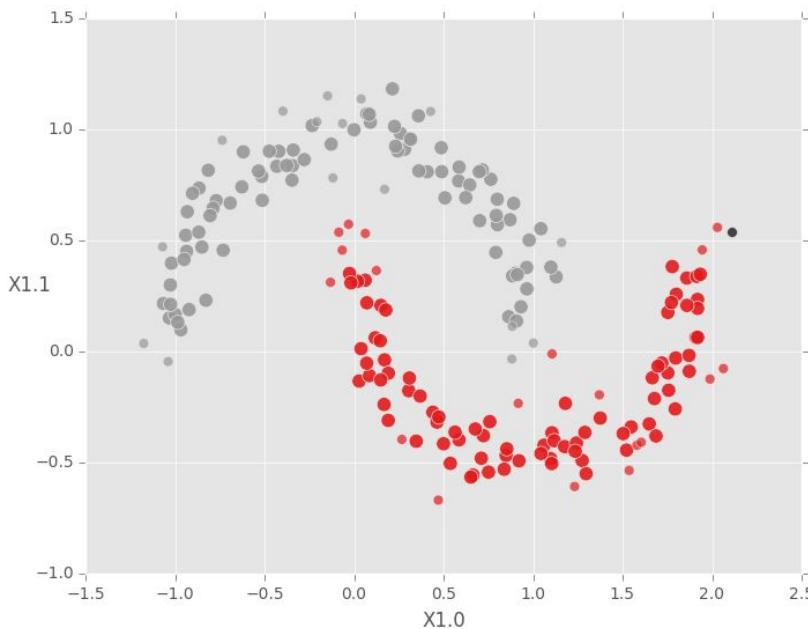
Exemplos



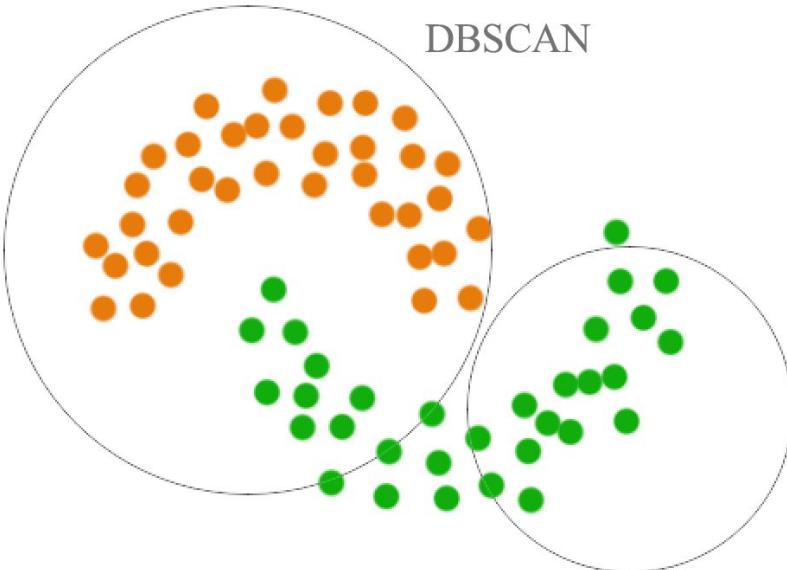
epsilon = 1.00
minPoints = 4

Fonte: https://cdn-images-1.medium.com/max/640/1*tc8UF-h0nQqUfLC8-0uInQ.gif

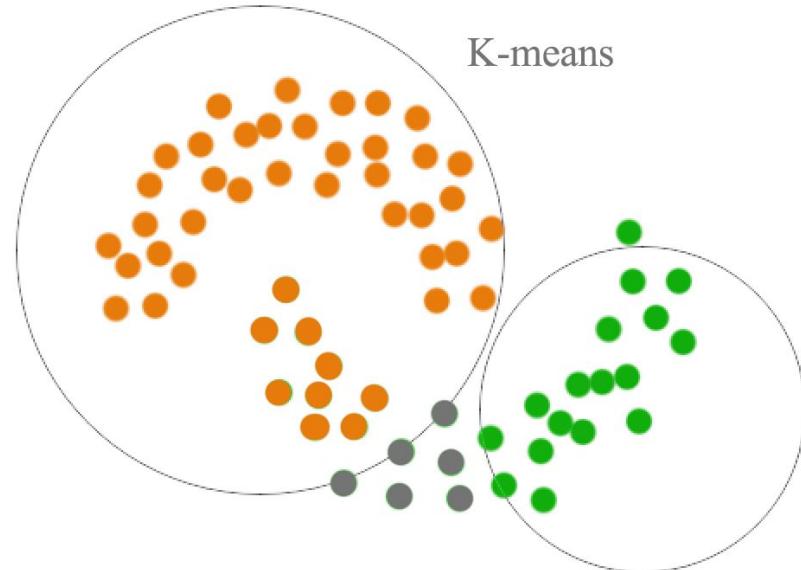
+ Exemplos



DBSCAN vs K-Means



DBSCAN



K-means

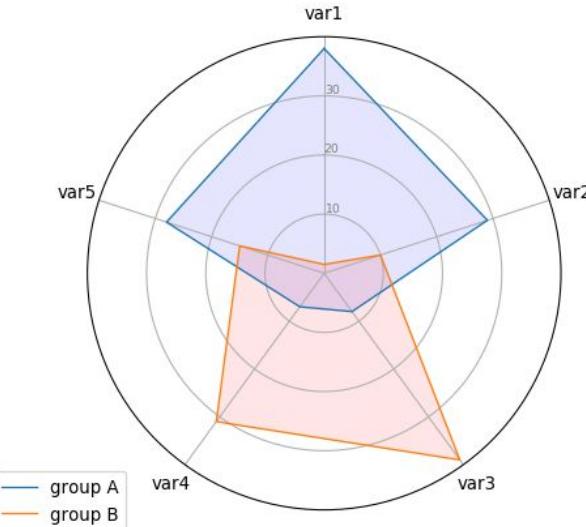


Como definir os hiperparâmetros?

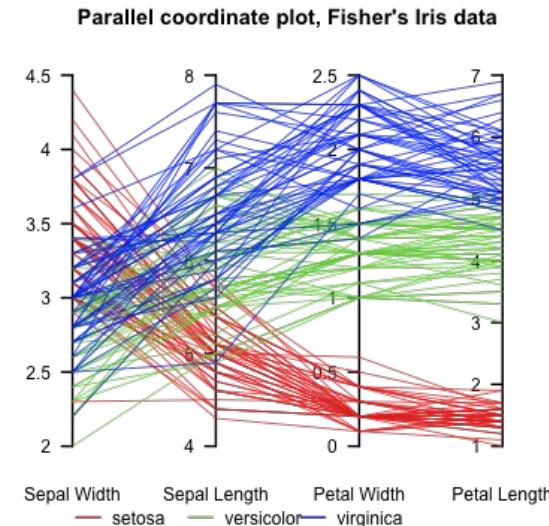
- ◊ Depende ...
- ◊ Um minPts baixo significa que ele criará mais clusters, inclusive a partir dos nós que deveriam ser ruído, logo, um valor muito pequeno não deve ser setado.
- ◊ Um raio grande juntamente com um baixo valor de minPts formará poucos clusters.
- ◊ Um raio pequeno com um baixo valor de minPts gerará muitos clusters
- ◊ Papers ...



Visualização



Radar chart



Parallel Coordinates



http://dontpad.com/kdd_uni7

Hands On

