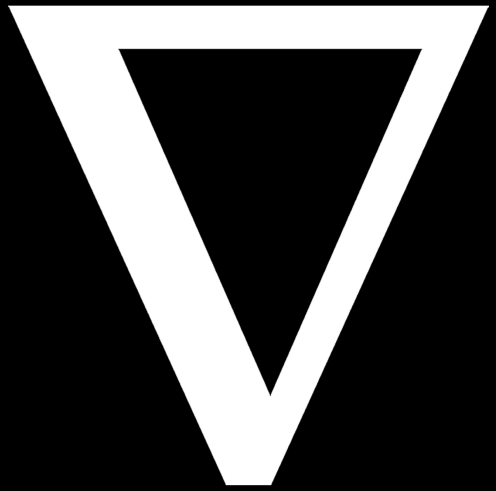
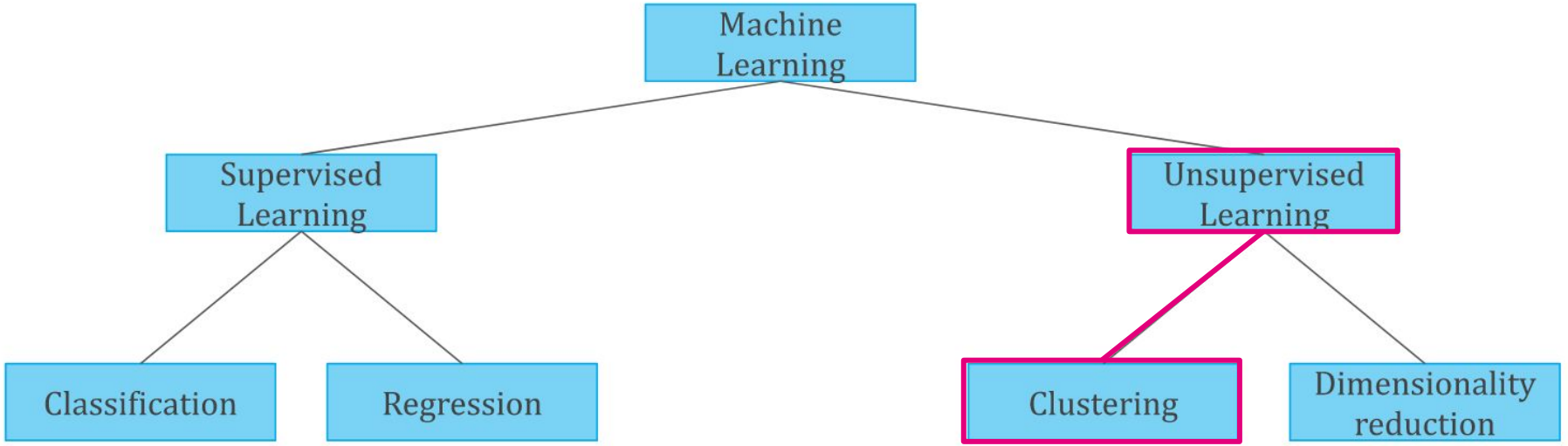


Clustering

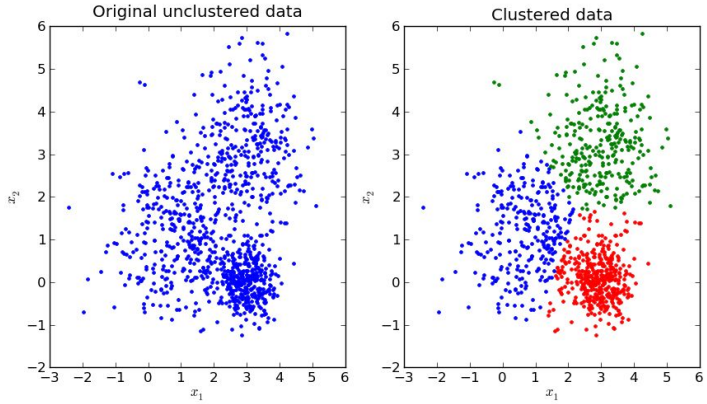
K-means e DBSCAN

Unsupervised Learning



Clustering

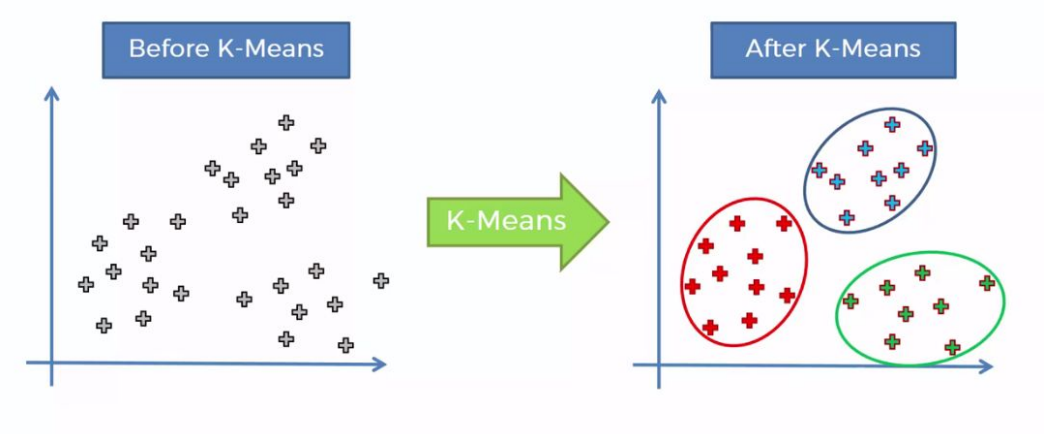
- Separar dados em um conjunto de grupos com base em semelhanças entre seus atributos
- Rotulagem de dados



K-means

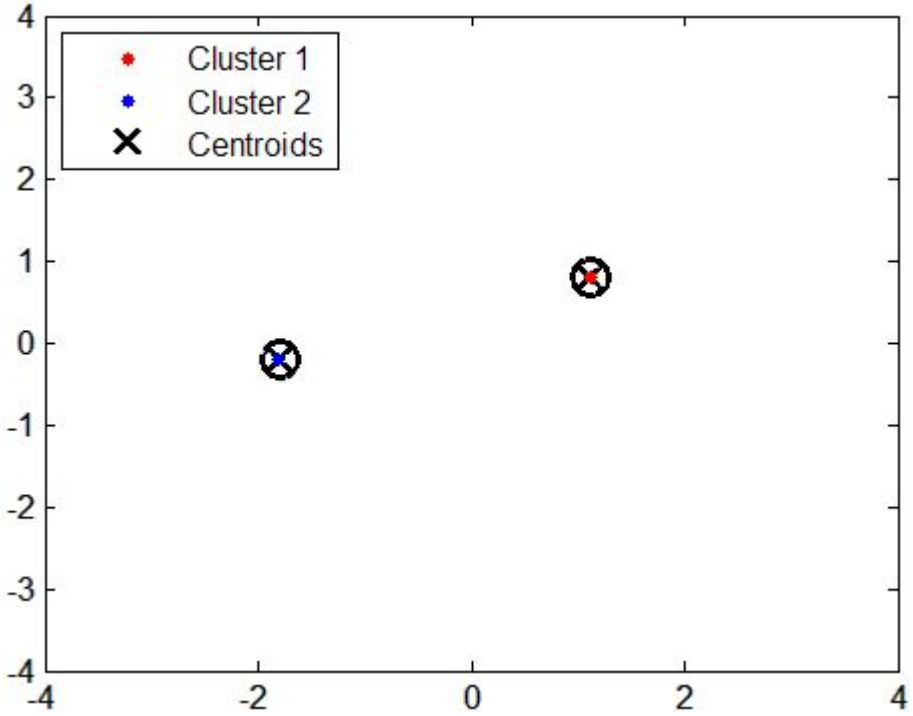
K-means

- Algoritmo de clustering
- Existem K clusters (grupos)
- Atribui a qual cluster cada dado pertence



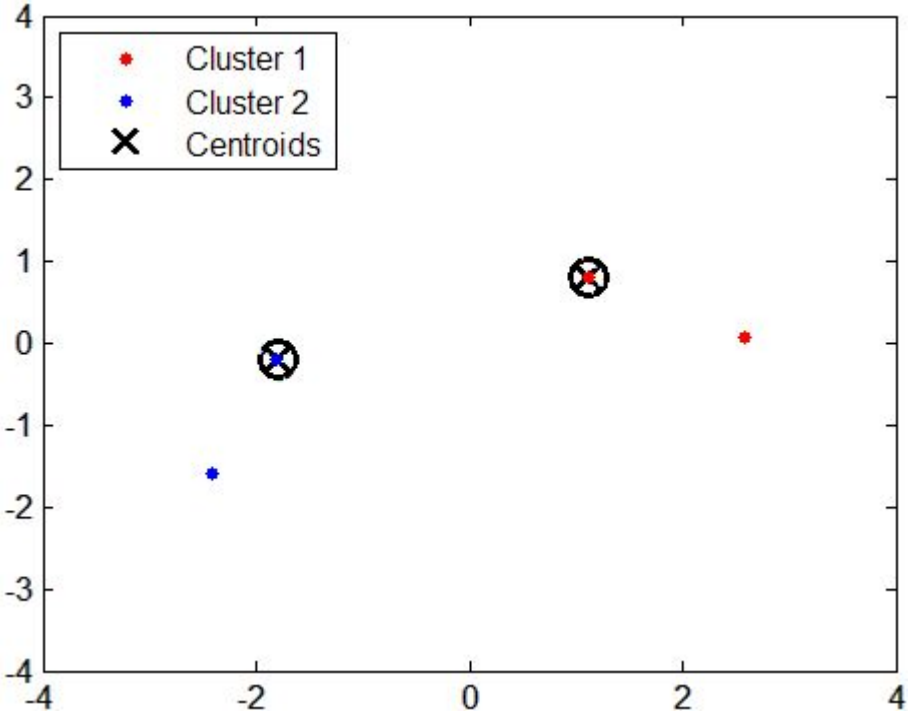
K-means(k=2): Algoritmo

1. São definidos centros de cada cluster



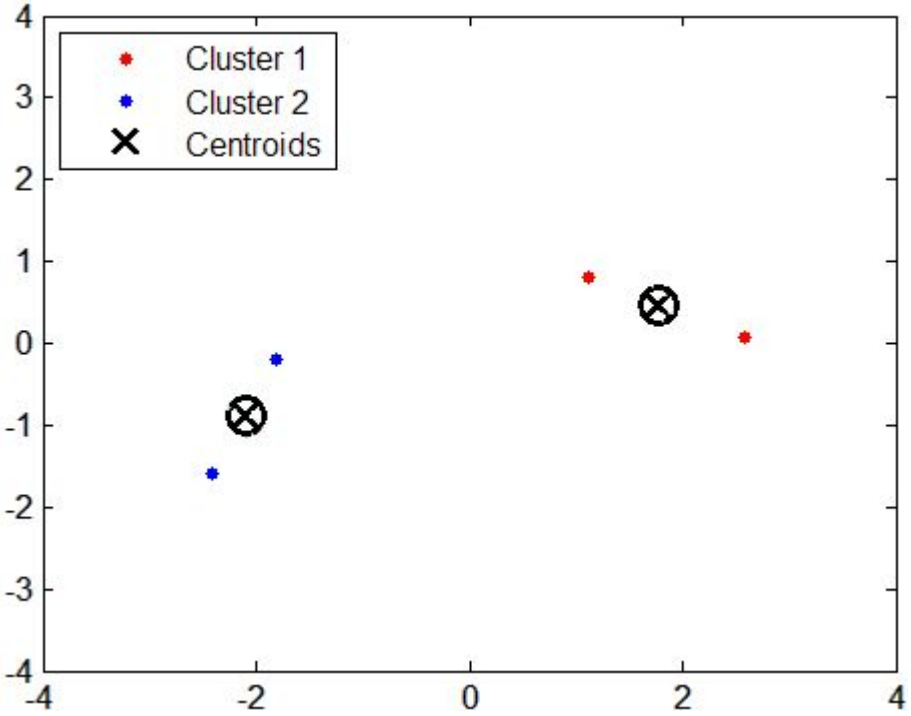
K-means(k=2): Algoritmo

1. São definidos centros de cada cluster
2. Dado é atribuído a um cluster: distância ao centro



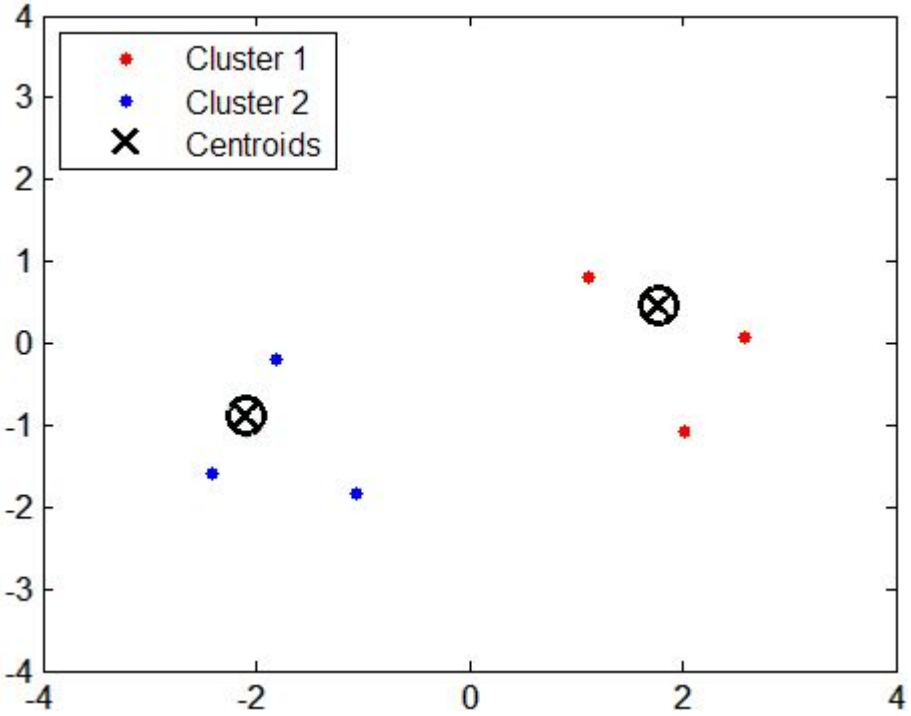
K-means(k=2): Algoritmo

1. São definidos centros de cada cluster
2. Dado é atribuído a um cluster: distância ao centro
3. O centro do cluster é recalculado: **Média** dos dados do cluster
4. Volta ao passo 2



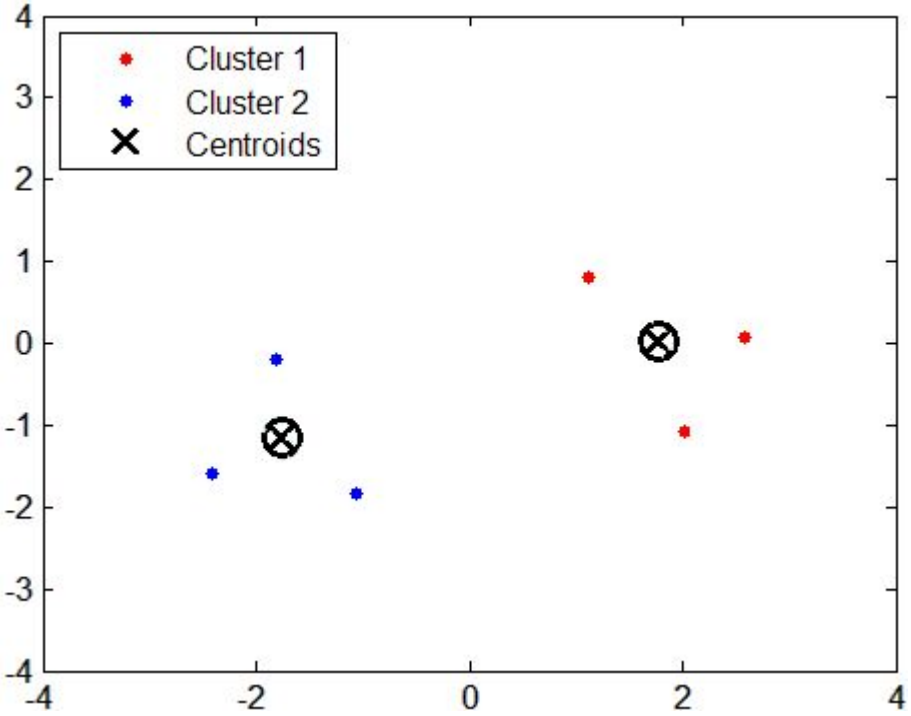
K-means(k=2): Algoritmo

1. São definidos centros de cada cluster
2. Dado é atribuído a um cluster: distância ao centro
3. O centro do cluster é recalculado: **Média** dos dados do cluster
4. Volta ao passo 2



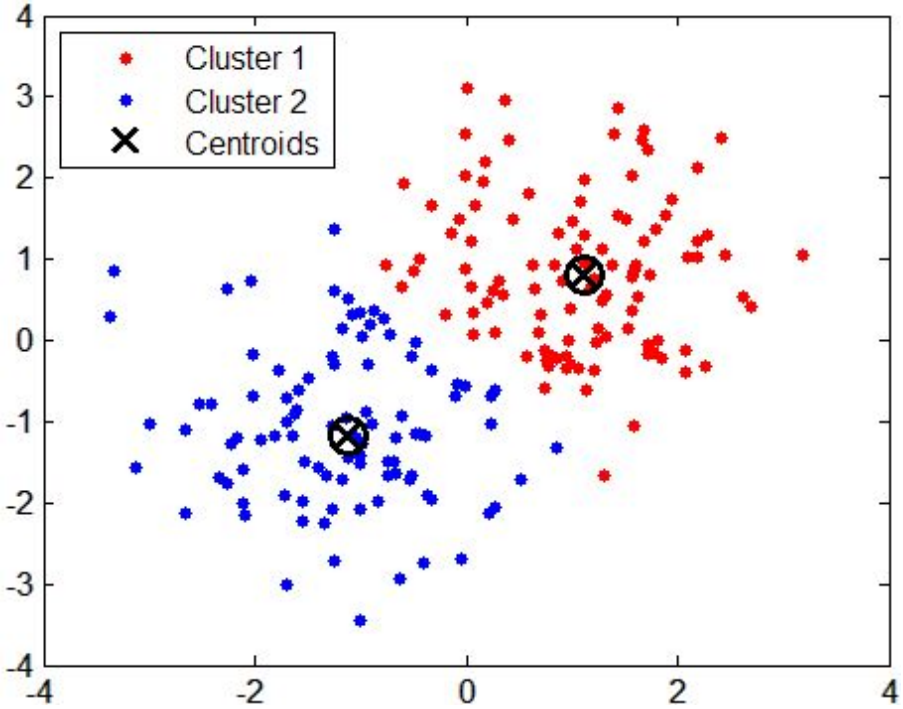
K-means: Algoritmo

1. São definidos centros de cada cluster
2. Dado é atribuído a um cluster: distância ao centro
3. O centro do cluster é recalculado: Média dos dados do cluster
4. Volta ao passo 2

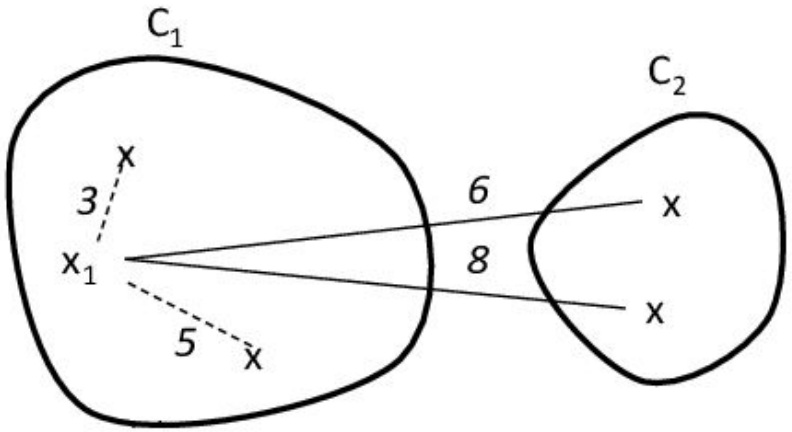


K-means(k=2): Algoritmo

1. São definidos centros de cada cluster
2. Dado é atribuído a um cluster: distância ao centro
3. O centro do cluster é recalculado: **Média** dos dados do cluster
4. Volta ao passo 2



Silhouette Coefficient


C1
C2
x
3
x1
5
x
6
8
x
x

Silhouette Coefficient

- Utilizado para avaliar a qualidade dos clusters formados
- Quão bem cada dado está agrupado com outros dados que são similares entre si
- É calculado com base em duas distâncias: Distância intra-cluster e Distância nearest-cluster

Silhouette Coefficient: Cálculo

Intra-Cluster (Cohesion)

Distância média entre um dado e todos os outros dados no mesmo cluster: a

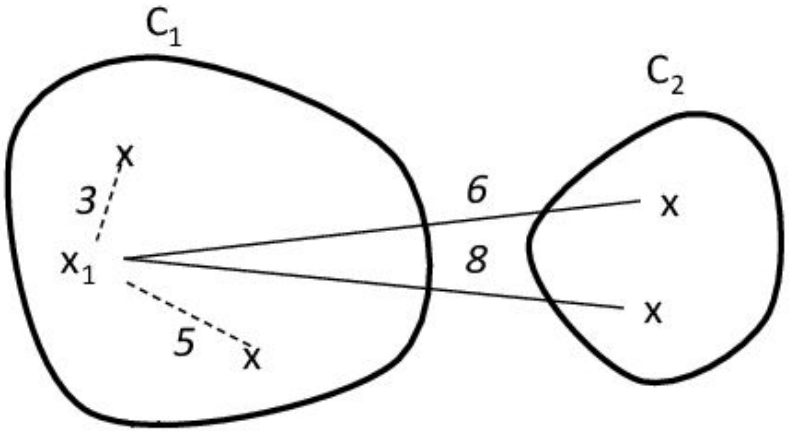
Nearest-Cluster (Separation)

Distância média entre um dado e todos os outros dados do cluster mais próximo: b

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

Silhouette Coefficient: Exemplo

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



$$a = \frac{3 + 5}{2} = 4$$

$$b = \frac{6 + 8}{2} = 7$$

$$S = \frac{7 - 4}{7} = \frac{3}{7}$$

Silhouette Coefficient: Interpretação

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

- S → -1: Dado foi definido no cluster errado
- S → 0: Não há distinção entre o cluster do dado e o cluster vizinho, clusters mal separados
- S → 1: Os dados do cluster estão bem próximos e bem separados dos clusters vizinhos

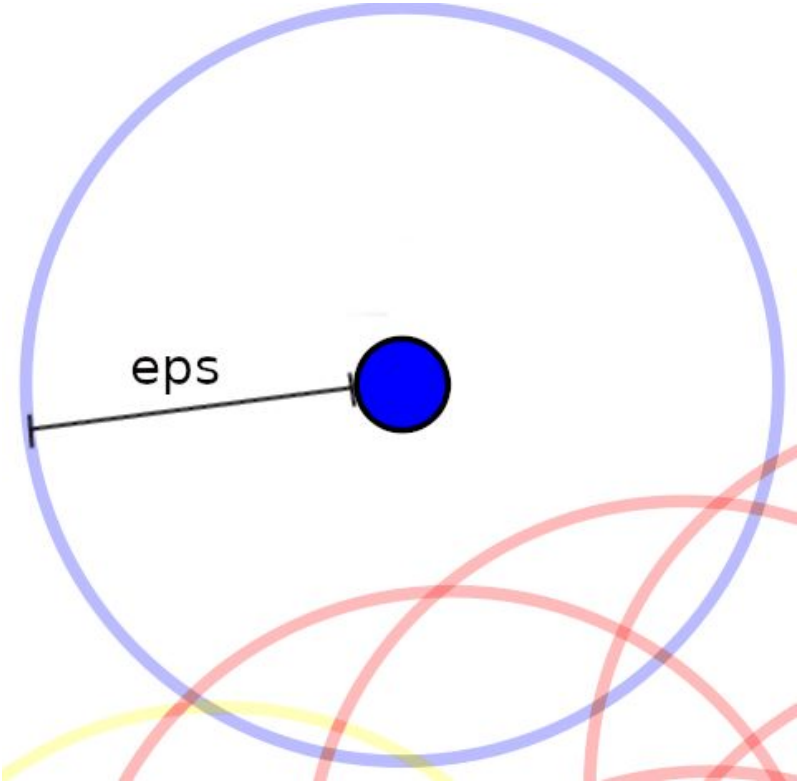
Silhouette Score: média dos coeficientes

demo

DBSCAN

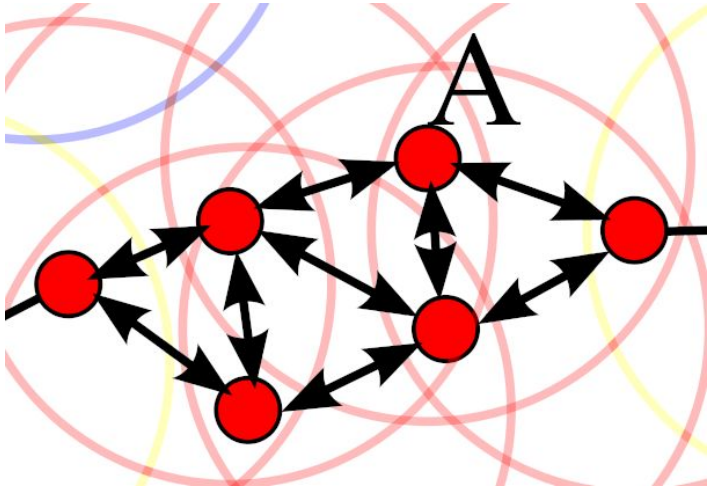
DBSCAN: Definições

- Epsilon (eps): Raio de um dado (ponto)
- Minimum Points (min Pts): Número mínimo de pontos dentro do raio para ser considerado como um ponto “core”



DBSCAN: Core Point

- São os pontos que formam o cluster
- Número de pontos dentro de eps é maior que min_Pts

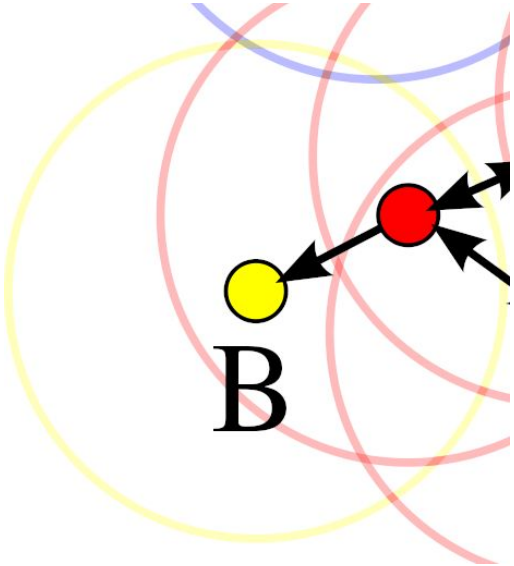


min_Pts = 4

DBSCAN: Border Point

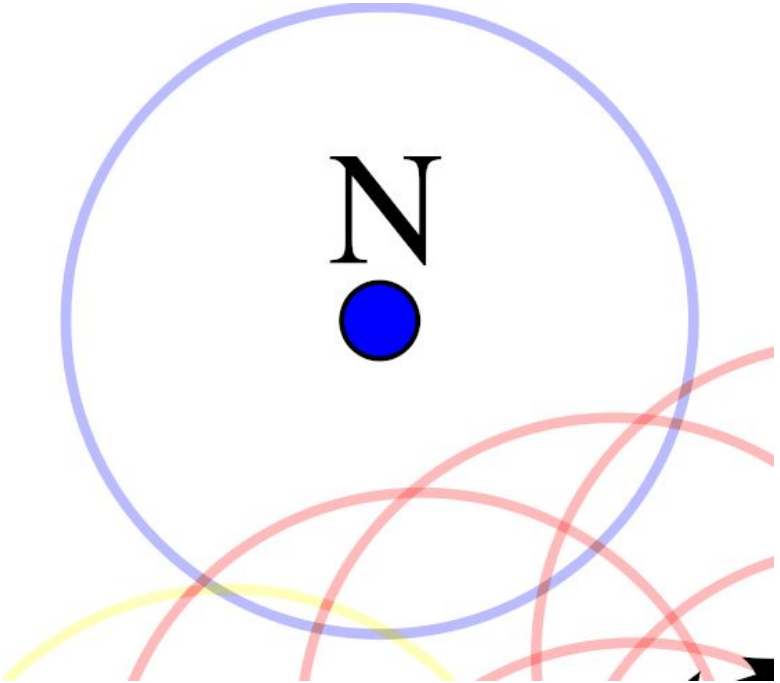
- Ainda fazem parte do cluster
- Há menos que min_Pts dentro de eps, porém há algum outro ponto

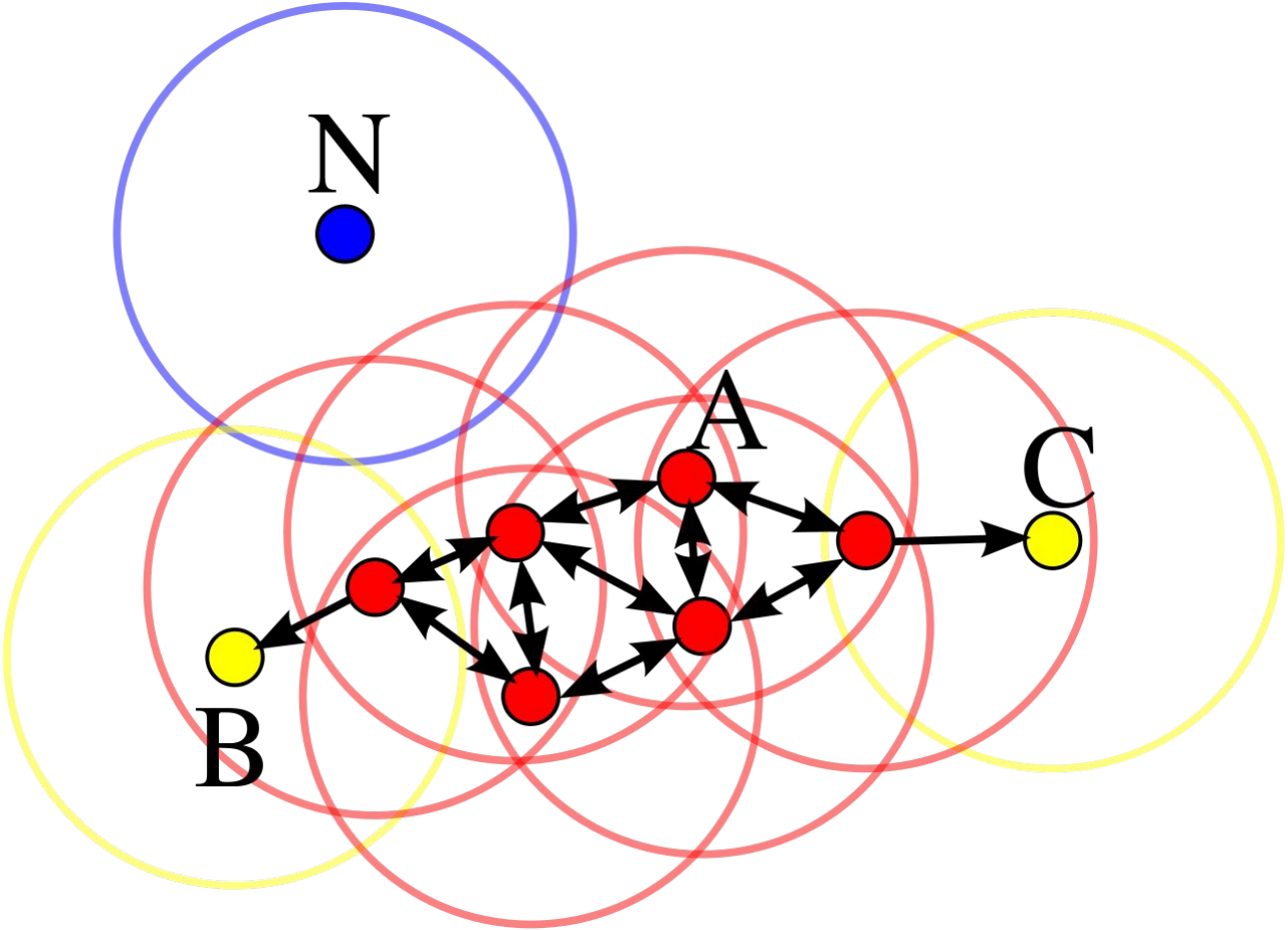
min_Pts = 4



DBSCAN: Noise Point

- Não fazem parte do cluster
- Outliers
- Não há nenhum outro ponto no raio de eps




N
A
C
B

DBSCAN: Vantagens e Desvantagens

Desvantagens:

- Não funciona bem para datasets com densidade variável
- Depende dos hiperparâmetros epsilon e minimum points

Vantagens:

- Lida bem com datasets que tenham bastante ruído
- Consegue identificar facilmente os outliers
- Clusters podem assumir uma forma irregular, ao contrário do K-Means, em que os clusters têm formatos esféricos

DBSCAN

k-means

demo

dúvidas?