FACULDADE IBRATEC

Agrupamento de Texto

Prof. Marcelo Pita

Pós em Inteligência Artificial

Processamento de Linguagem Natural

Agrupamento

Agrupamento (*clustering*) consiste na tarefa de agrupar conjuntos de dados de acordo com medidas de similaridade ou distância.

Agrupamento

Técnicas de agrupamento são métodos usados para a construção de *grupos de objetos* com base nas semelhanças e diferenças entre os mesmos de tal maneira que os grupos obtidos são os mais homogêneos e o mais separados possíveis.

Objetos: abstração para todo tipo de estrutura

- Registros
- Documentos
- Grafos

Agrupamento

Para que serve?

- Segmentação de mercado
- Agrupamento de grafos em análise de redes sociais
- Agrupamento de coleções de documentos (não confundir com tópicos)
- Análise de formação de galáxias
- Identificação de grupos de segurados com um custo médio elevado de reembolso
- Identificação de grupos de habitação segundo o tipo, valor e localização geográfica

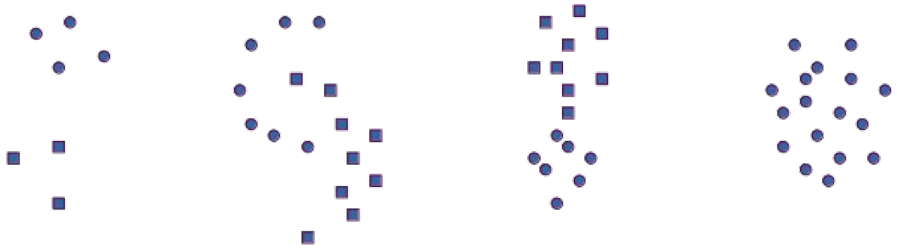
Agrupamento

Um bom método de agrupamento fornece grupos de alta qualidade com:

- Alta similaridade intra-grupo
- Baixa similaridade inter-grupo

A qualidade do resultado de um agrupamento depende tanto da medida de similaridade usada pelo método como da sua implementação.

Agrupamento



a) Grupos coesos e isolados

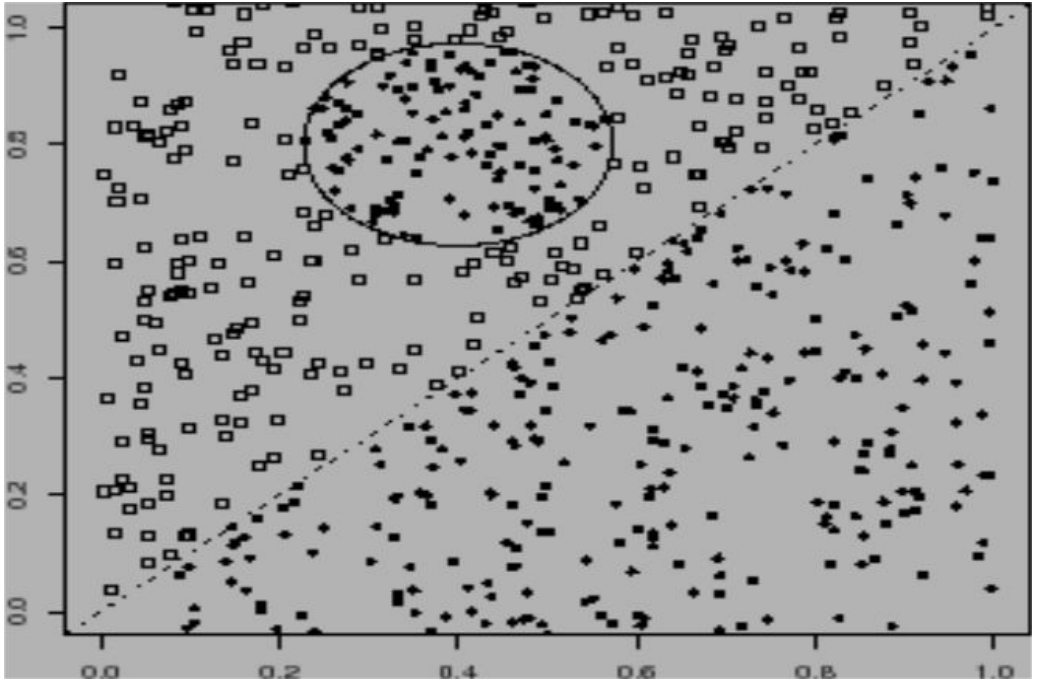
b) Grupos isolados mas não coesos

c) Grupos coesos com vários pontos intermediários

d) Não existência de grupos “naturais”

Agrupamento

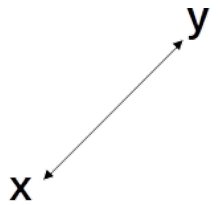
Dependendo das características dos dados o processo de agrupamento pode ser difícil.



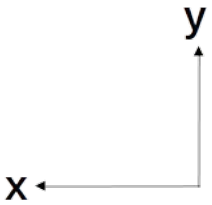
Medidas de similaridade

Uma forma de determinar se dois registros são próximos é através de *medidas de similaridade e distância*.

Distância Euclidiana $$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Distância de Manhatan $$d = \sum_{i=1}^{n} |x_i - y_i|$$



Classes de métodos de agrupamento

Métodos baseados em centróide: Encontrar k centros de grupos e atribuir cada objetos ao centro mais próximo. Exemplos: k-means; k-medoids; k-medians.

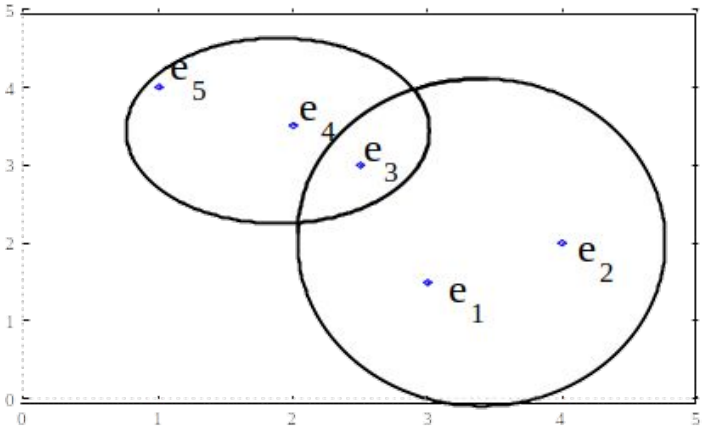
Métodos de Densidade: Grupos são regiões mais densas na distribuição dos dados. Exemplos: DBSCAN; OPTICS.

Métodos baseados em distribuição: Supõe-se uma distribuição de probabilidade para cada grupo. Exemplo: GMM (Gaussian Mixture Models)

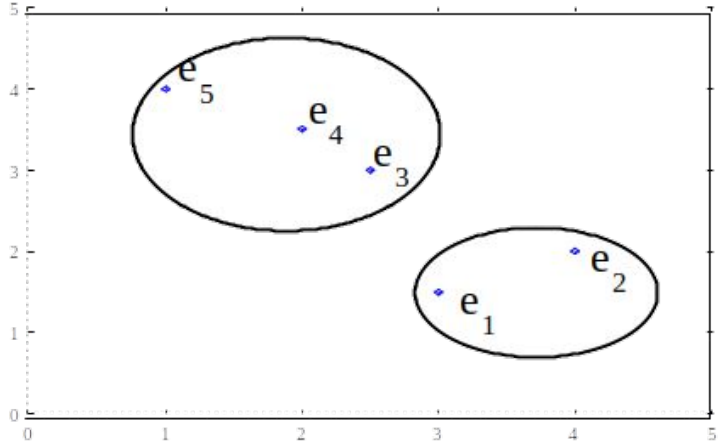
Métodos hierárquicos ou baseados em conectividade: A ideia é a de que um objeto está mais relacionado com objetos mais próximos do que com mais distantes. Produzem uma hierarquia de grupos.

Classes de métodos de agrupamento

Cobertura

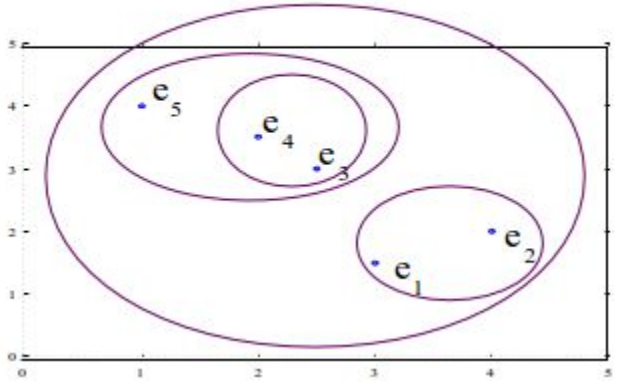


Partição

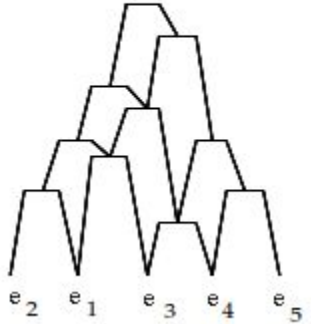


Classes de métodos de agrupamento

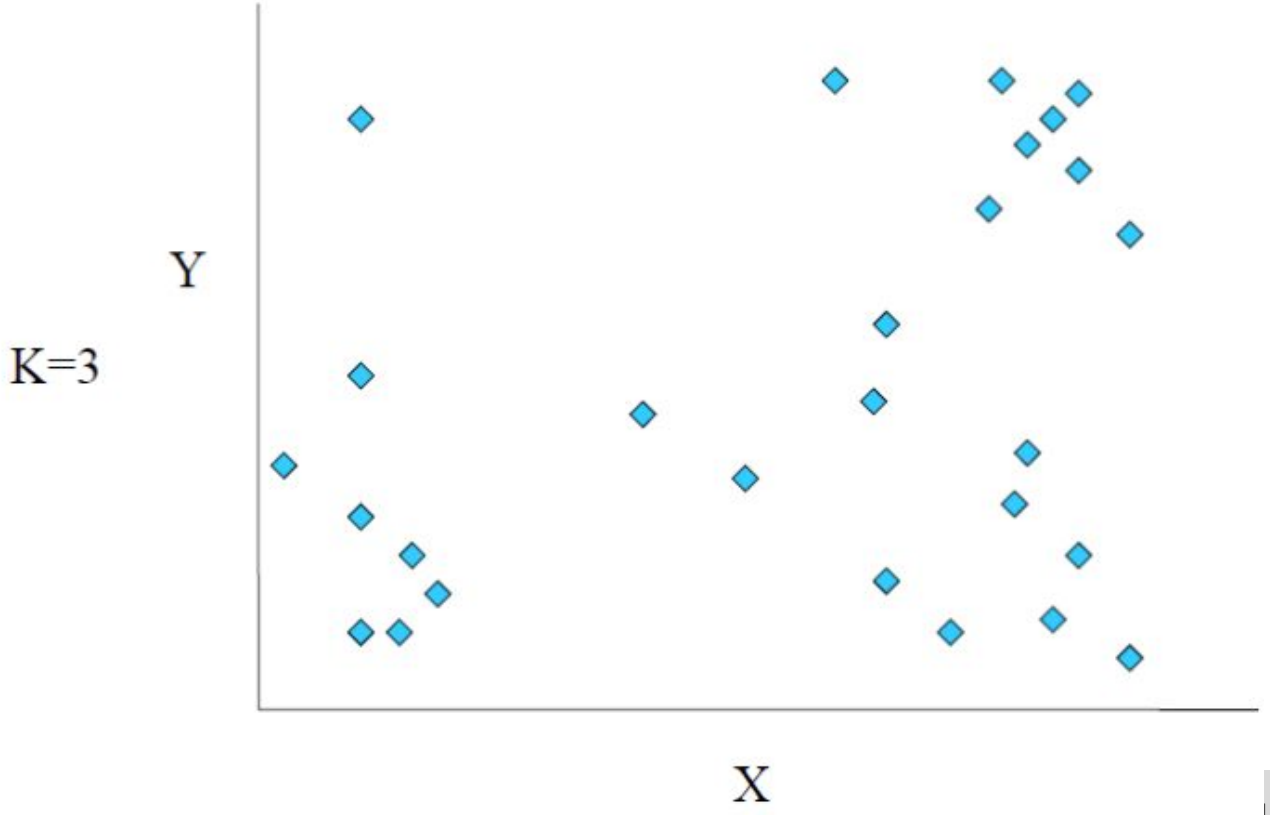
Hierarquia



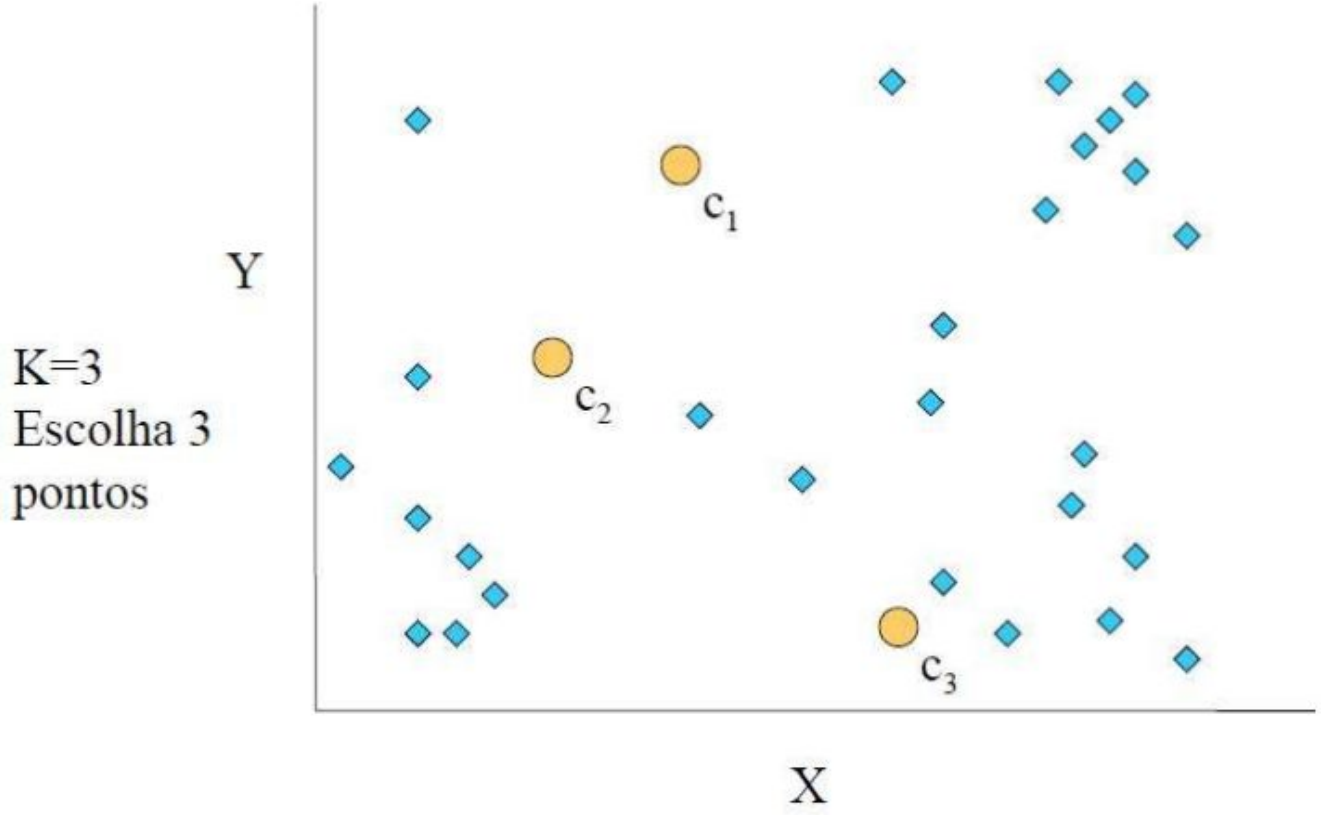
Piramide



Algoritmo K-Means

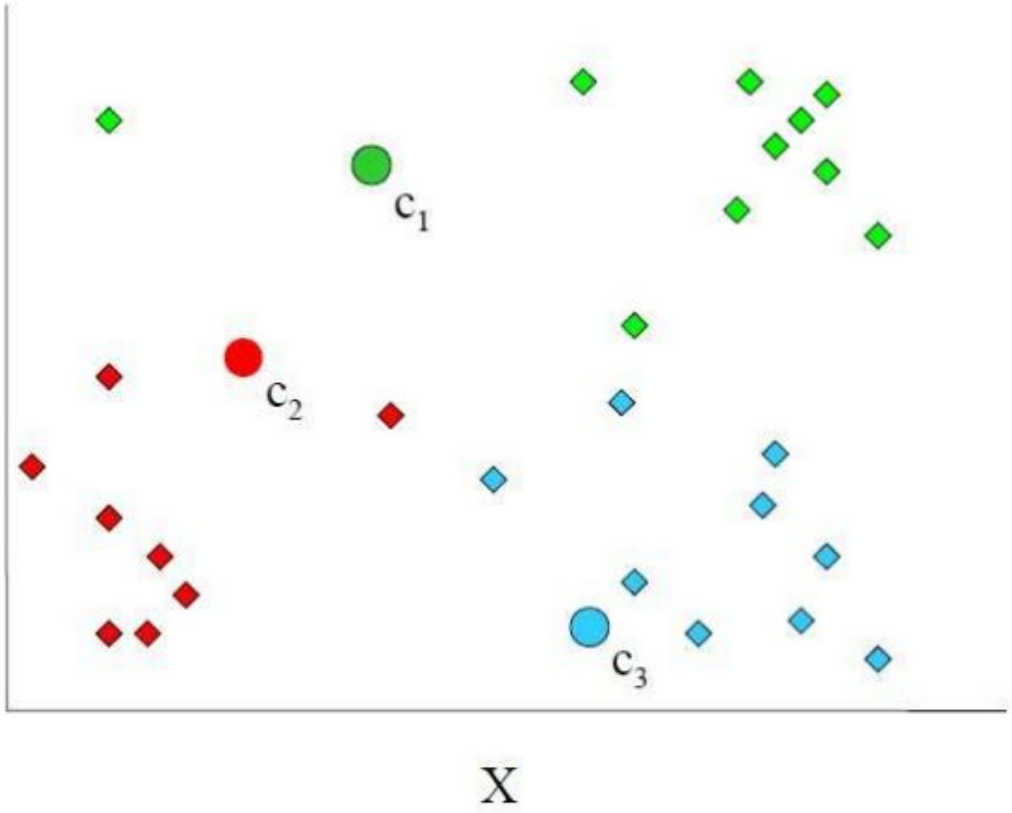


Algoritmo K-Means

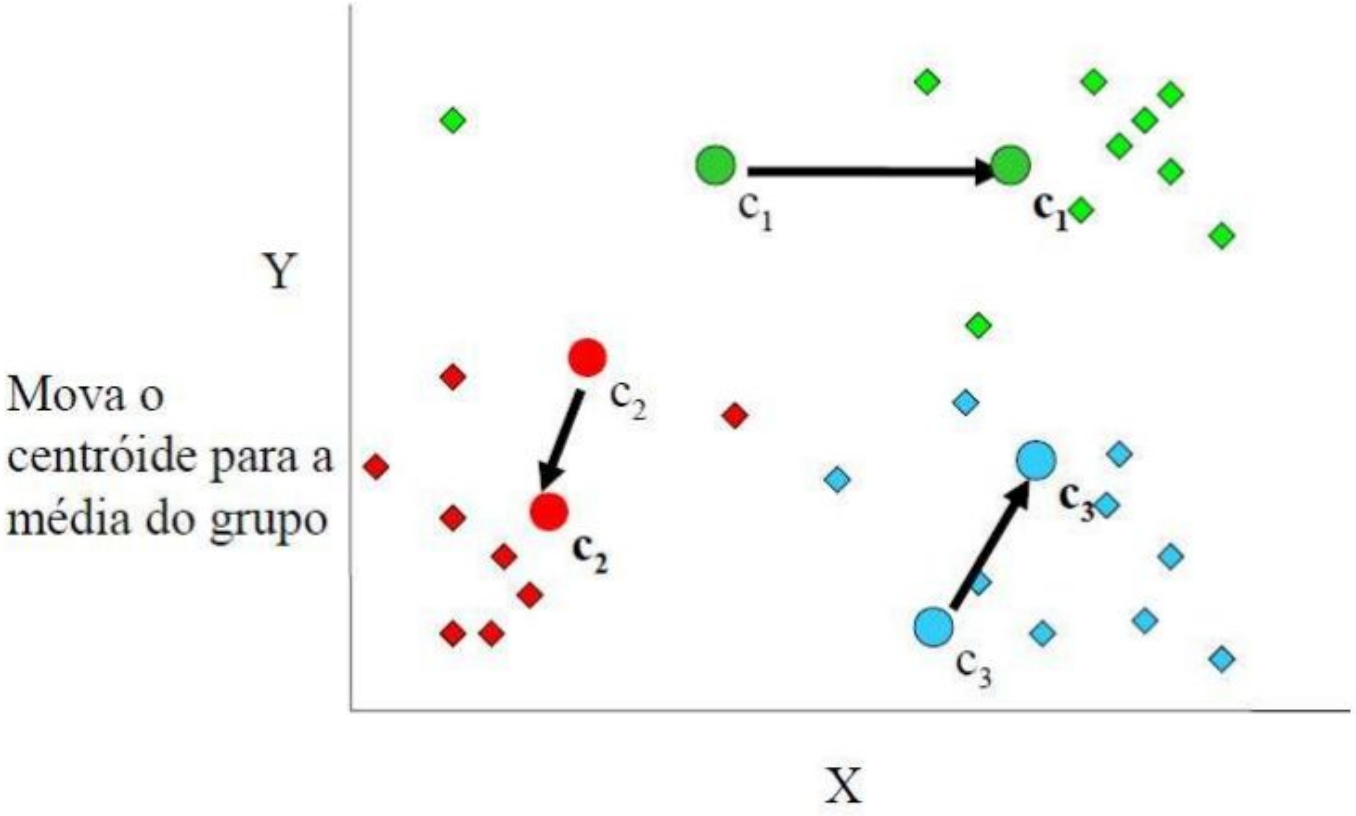


Algoritmo K-Means

Cada ponto é assinalado ao centróide mais próximo.

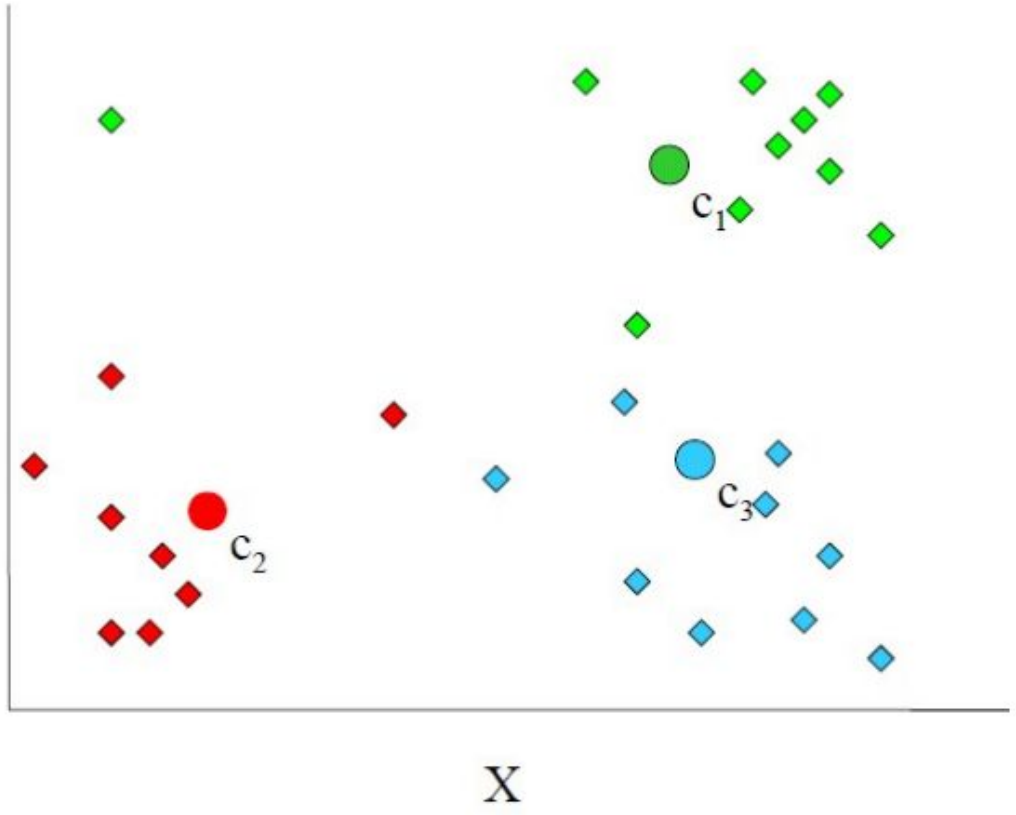


Algoritmo K-Means

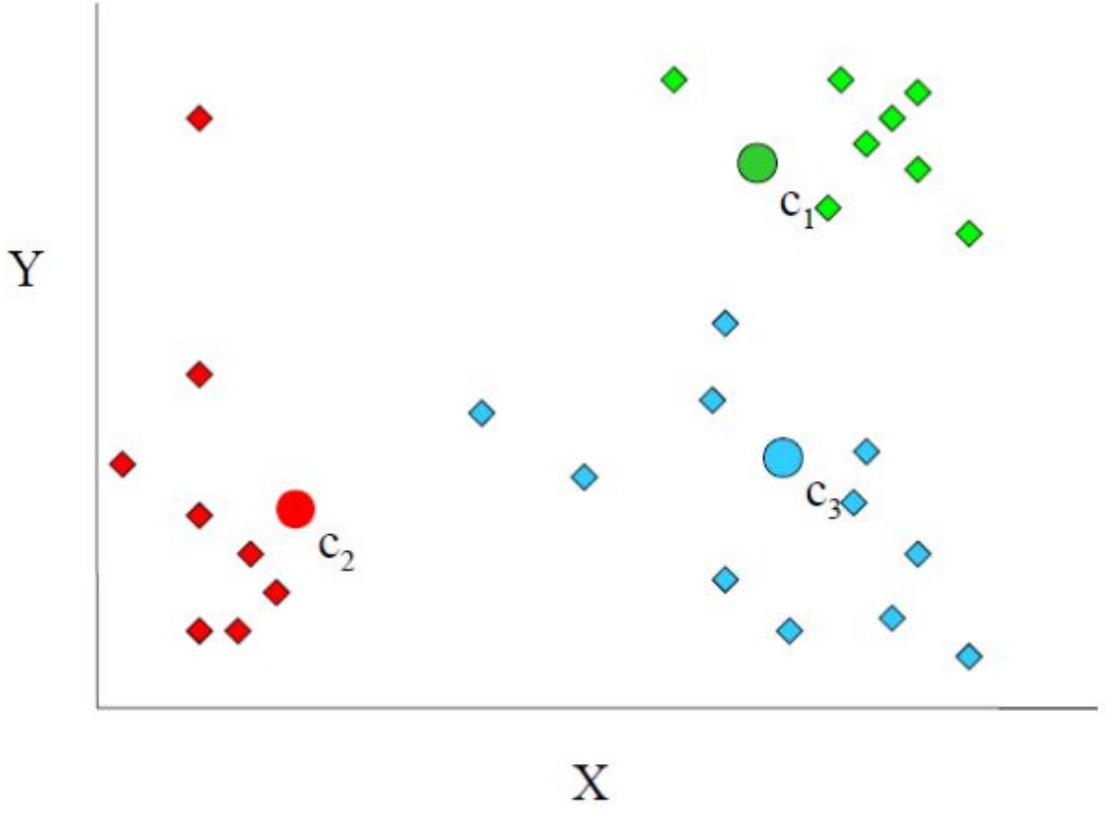


Algoritmo K-Means

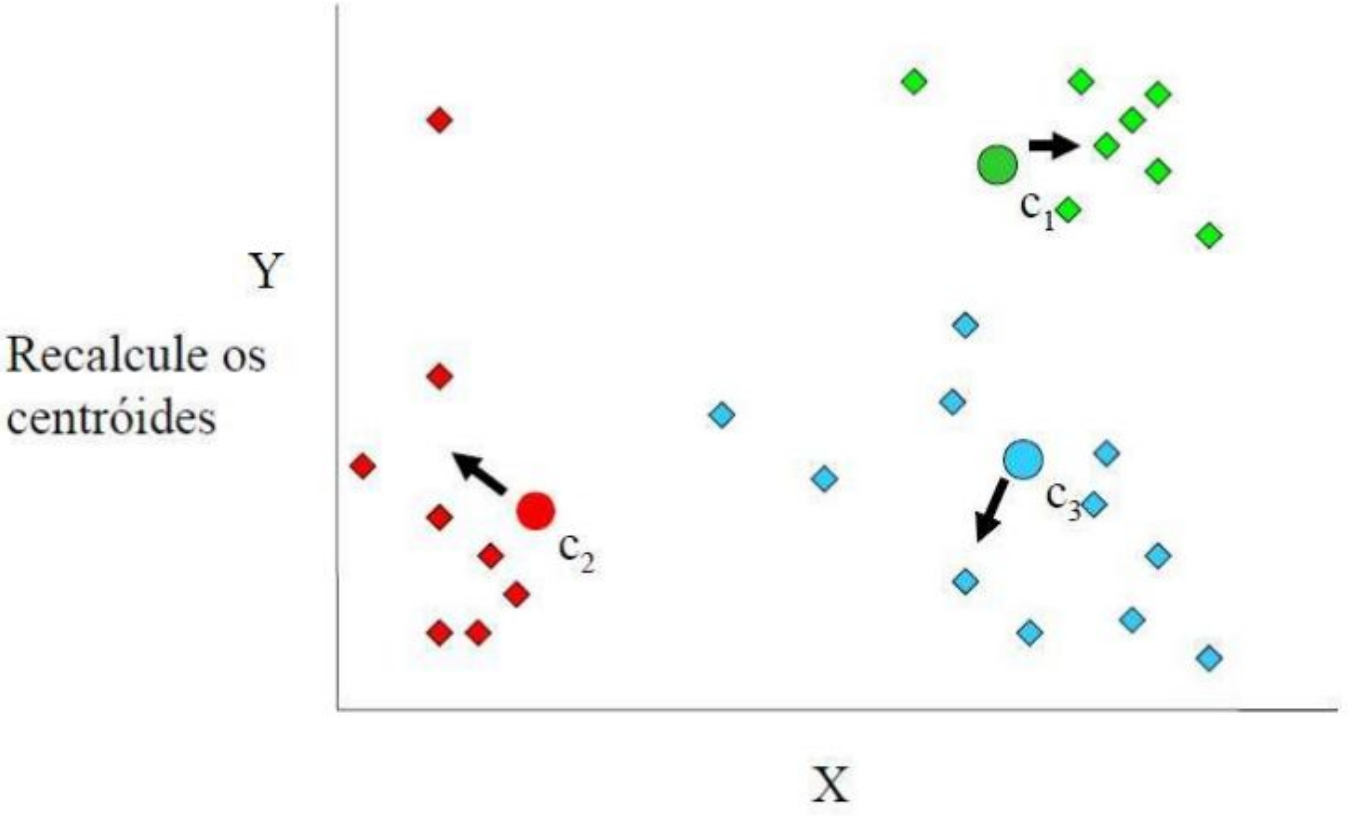
Reassinale os pontos para o centróide mais próximo



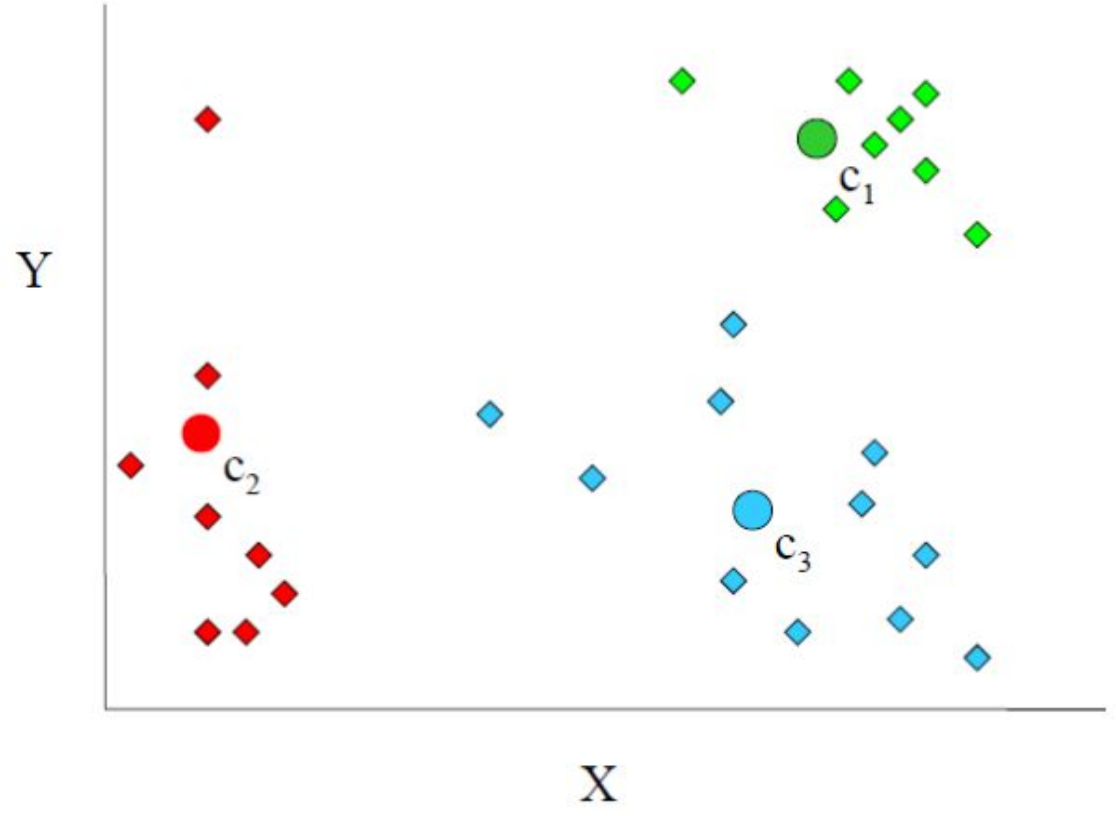
Algoritmo K-Means



Algoritmo K-Means



Algoritmo K-Means



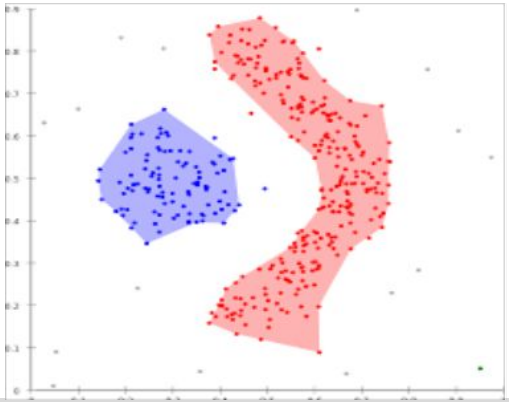
Algoritmo K-Means

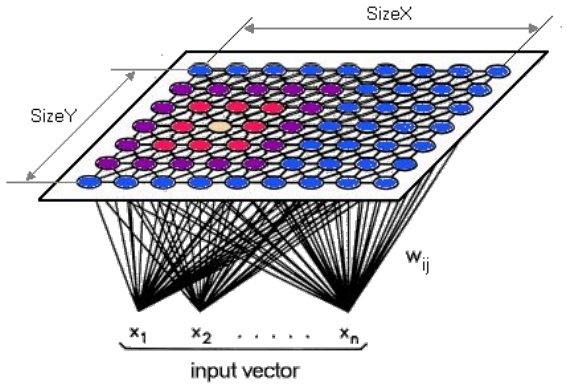
Nem sempre será ótimo, pois depende da atribuição inicial dos centróides.

Outros algoritmos

Existem muitos outros algoritmos:

- DBSCAN - Baseado em densidade
- Mapas auto-organizáveis (mapas de Kohonem, redes SOM)





Agrupamento de texto

Aplicação das técnicas de agrupamento conhecidas para o domínio de texto.

Uso de representações adequadas:

- TF, TF-IDF, *word embeddings*

Uso de métricas de similaridade adequadas:

- Similaridade do cosseno, Euclidiana.

Agrupamento de texto

<Prática>