Pontifícia Universidade Católica do Paraná

Inteligência Computacional:
Agrupamentos

Aprendizagem Não-Supervisionada

- Quando os dados ou amostras de um problema não estão rotulados.
- Tarefa principal: Agrupamento
- Descobrir grupos (clusters) – instâncias similares segundo algum critério.

Aprendizagem Não-Supervisionada

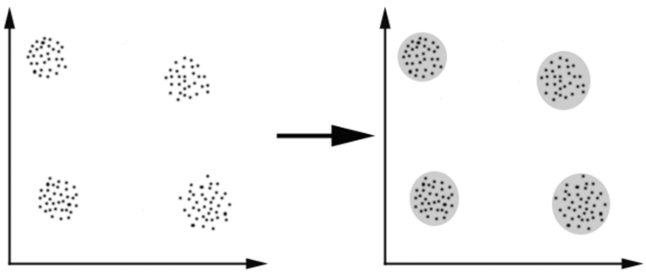
- Motivação:
  - Rotular bases de dados é uma tarefa de alto custo
  - Não é raro não se ter conhecimento das classes do problema.
  - Mineração e descoberta de conhecimento.

Aprendizagem Não-Supervisionada

- Um passo antes da criação de um classificador:
  - Dada uma base de dados não rotulada, pode-se utilizar a aprendizagem não-supervisionada para fazer uma pré-classificação, e então treinar um classificador de maneira supervisionada.

Agrupamento (Clustering)

- Organização de objetos em grupos (clusters) segundo algum critério de similaridade.



Quatro grupos (clusters)

Cluster

- Uma coleção de objetos que são similares entre si, e diferentes dos objetos pertencentes a outros clusters.
- Isso requer uma medida de similaridade.
- Usualmente uma *distância.*
  - *Distance-based Clustering*

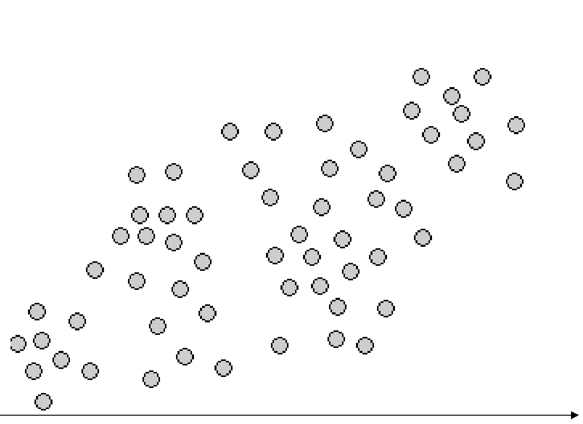
k-Means Clustering

- É a técnica mais simples de aprendizagem não supervisionada.
- Consiste em fixar *k* centróides (de maneira aleatória), um para cada grupo (cluster).
- Associar cada indivíduo ao centróide mais próximo.
- Recalcular os centróides com base nos indivíduos classificados.

Algoritmo *k-Means*

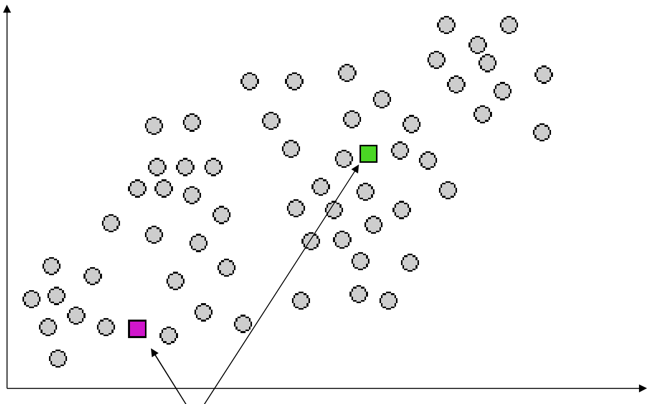
1. Determinar os centróides
2. Atribuir a cada objeto do grupo o centróide mais próximo.
3. Após atribuir um centróide a cada objeto, recalcular os centróides.
4. Repetir os passos 2 e 3 até que os centróides não sejam modificados.

k-Means – Um Exemplo



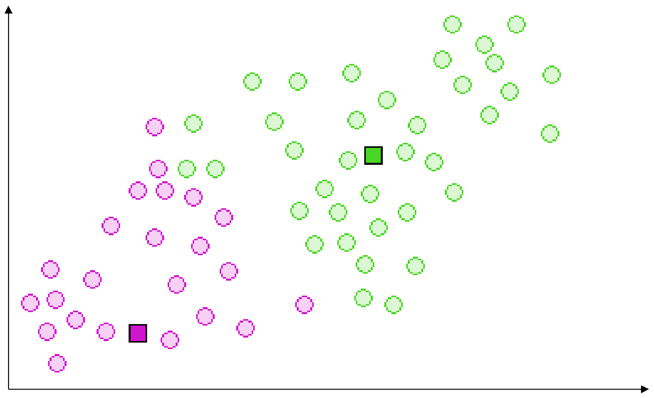
Objetos em um plano 2D

k-Means – Um Exemplo



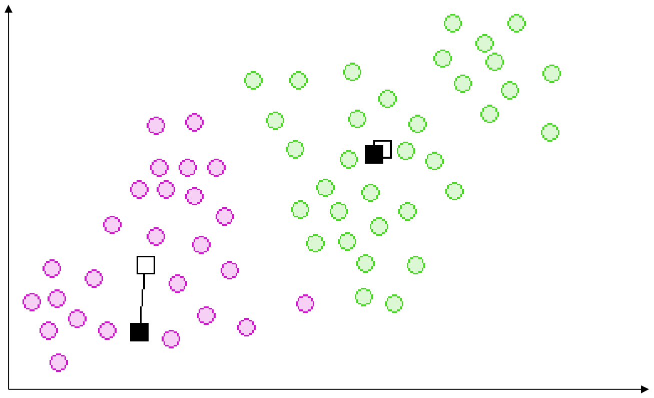
Passo 1:Centróides inseridos aleatoriamente

k-Means – Um Exemplo



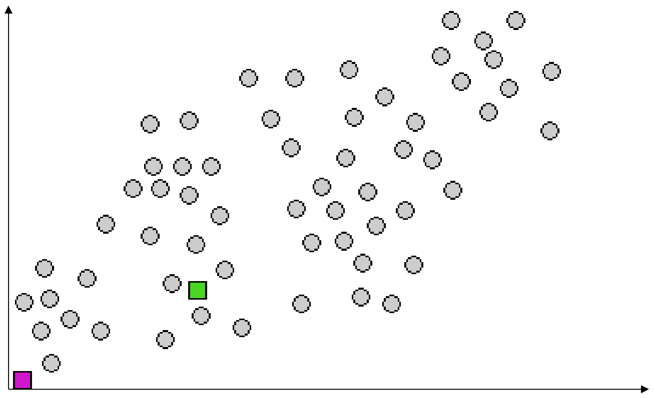
Passo 2: Atribuir a cada objeto o centróide mais próximo

k-Means – Um Exemplo



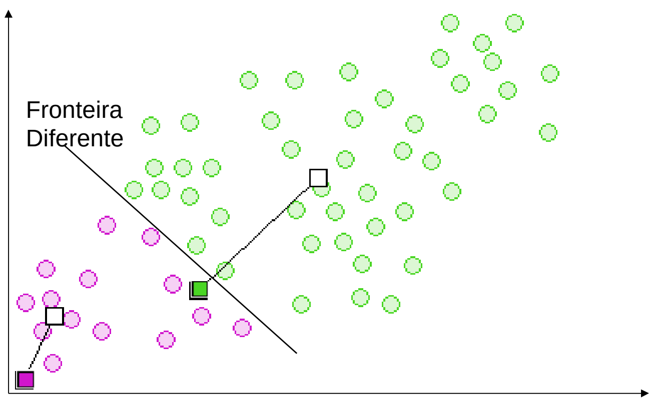
Passo 3: Recalcular os centróides

k-Means – Um Exemplo



Impacto da inicialização aleatória.

k-Means – Um Exemplo

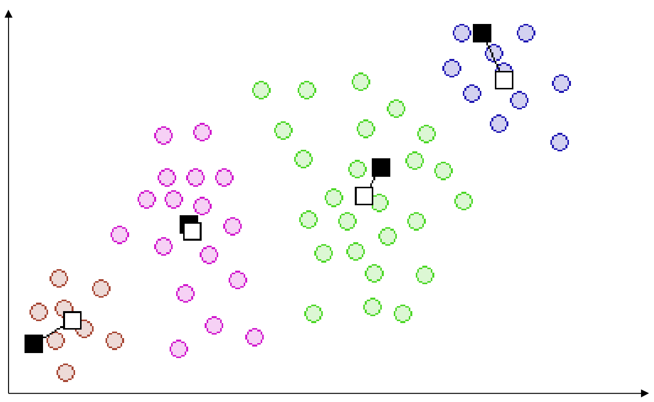


Impacto da inicialização aleatória

k-Means – Inicialização

- Importância da inicialização.
- Quando se tem noção dos centróides, pode-se melhorar a convergência do algoritmo.
- Execução do algoritmo várias vezes, permite reduzir impacto da inicialização aleatória.

k-Means – Um Exemplo

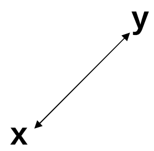


4 Centróides

Calculando Distâncias

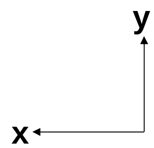
- Distância Euclidiana

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



- Manhattan (City Block)

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



Calculando Distâncias

- Minkowski
  - Parâmetro r
    - r = 2, distância Euclidiana
    - r = 1, City Block

$$d = \left(\sum_{i=1}^{n} (x_i - y_i)^r \right)^{1/r}$$

Calculando Distâncias

- Mahalanobis
  - Leva em consideração as variações estatísticas dos pontos. Por exemplo, se x e y são dois pontos da mesma distribuição, com matriz de covariância C, a distância é dada pela equação

$$d = \left((x-y)' C^{-1} (x-y)\right)^{\frac{1}{2}}$$

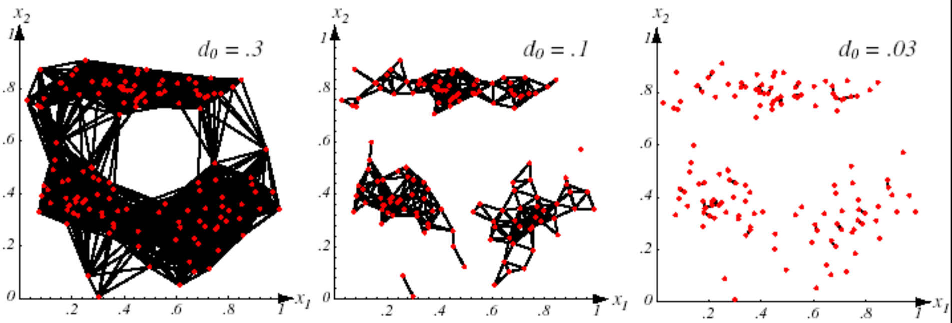
  - Se a matriz C for uma matriz identidade, essa distância é igual a distância Euclidiana.

A Importância das Medidas de Distâncias

- Suponha que dois exemplos pertencem ao mesmo cluster se a distância Euclidiana entre eles for menor que *d.*
- É obvio que a escolha de *d* é importante.
- Se *d* for muito grande, provavelmente teremos um único cluster, se for muito pequeno, vários clusters.

A Importância das Medidas de Distâncias

- Nesse caso, estamos definido *d* e não *k.*



Critérios de Otimização

- Até agora discutimos somente como medir a similaridade.
- Um outro aspecto importante em *clustering* é o critério a ser otimizado.
- Considere um conjunto $D = \{x_1, ..., x_n\}$ composto de *n* exemplos, e que deve ser dividido em *c* sub-conjuntos disjuntos $D_1, ..., D_c$.
- Cada sub-conjunto representa um *cluster*.

Critérios de Otimização

- O problema consiste em encontrar os *clusters* que minimizam/maximizam um dado critério.
- Alguns critérios de otimização:
  - Soma dos Erros Quadrados.
  - Critérios de Dispersão

Soma dos Erros Quadrados

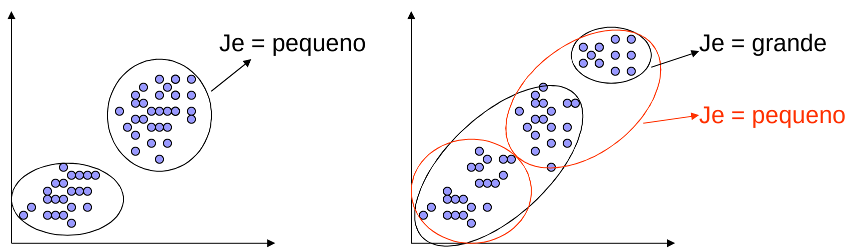
- É o mais simples e usado critério de otimização em *clustering.*
- Seja n_i o número de exemplos no cluster D_i e $\mathbf{m_i}$ a média desse exemplos

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

- A soma dos erros quadrados é definida

$$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} (x - m_i)^2$$

Soma dos Erros Quadrados



Adequado nesses casos
- Separação natural

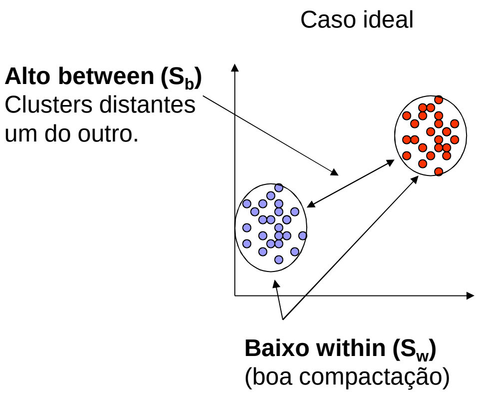
Não é muito adequado para dados mais dispersos.
Outliers podem afetar bastante os vetores médios **m**

Critérios de Dispersão

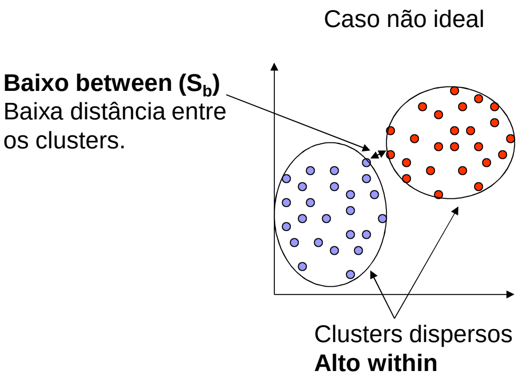
- Vetor médio do cluster i $m_i = \frac{1}{n_i} \sum_{x \in D_i} x$
- Vetor médio total $\mathrm{m} = \frac{1}{n} \sum_{D} x$
- Dispersão do cluster i $S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^t$
- Within-cluster $S_w = \sum_{i=1}^{c} S_i$
- Between-cluster $S_B = \sum_{i=1}^{c} n_i (m_i - \mathrm{m})(m_i - \mathrm{m})^t$

Critérios de Dispersão

- Relação Within-Between

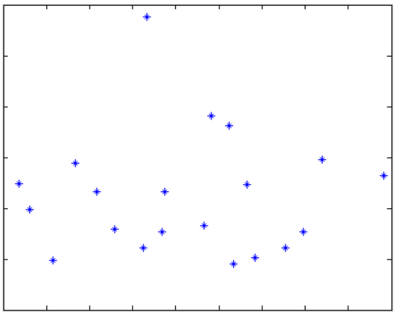


Critérios de Dispersão

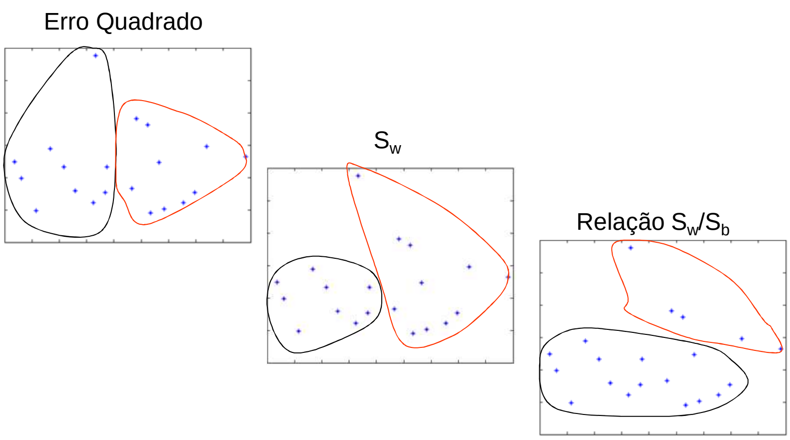


Critérios de Dispersão

- Podemos entender melhor os critérios de dispersão analisando o seguinte exemplo:



Diferentes clusters para c=2 usando diferentes critérios de otimização



Algumas Aplicações de *Clustering*

- Marketing: Encontrar grupos de consumidores com comportamento similares
- Biologia: Classificar grupos de plantas e animais.
- Bibliotecas: Organização de livros.
- Administração: Organização de cidades, classificando casas de acordo com suas características.
- WWW: Classificação de conteúdos.

Problemas

- Vetores de característica muito grandes: tempo de processamento elevado.
- Definição da melhor medida de distância: Depende do problema. As vezes é difícil, especialmente quando se trabalha com grandes dimensões.
- O resultado do *clustering* pode ser interpretado de diferentes maneiras.

k-Means - Simulação

- Um *applet* java para a simulação do *k-Means* pode ser encontrado na seguinte URL:
  http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html

Principais Técnicas

- K-means
- X-Means: K-means, onde K é definido automaticamente. Usa BIC (Bayesian Information Criterion).
- Fuzzy C-means: usa noção de pertinência. Uma instância pode pertencer a mais de um cluster.
- Hirárquico: organiza os grupos em uma estrutura hierárquica.
- Mixture of Gaussians: baseado em modelo. EM (Expectation Maximization)