Machine Learning

Dia 10 - Aprendizado não-supervisionado

ImageU - Grupo de Pesquisa em Machine Learning e Visão Computacional
https://imageu.github.io/

Curso de Verão 2022

Instituto de Matemática e Estatística - IME USP

Programa

1. Aprendizado Não-Supervisionado

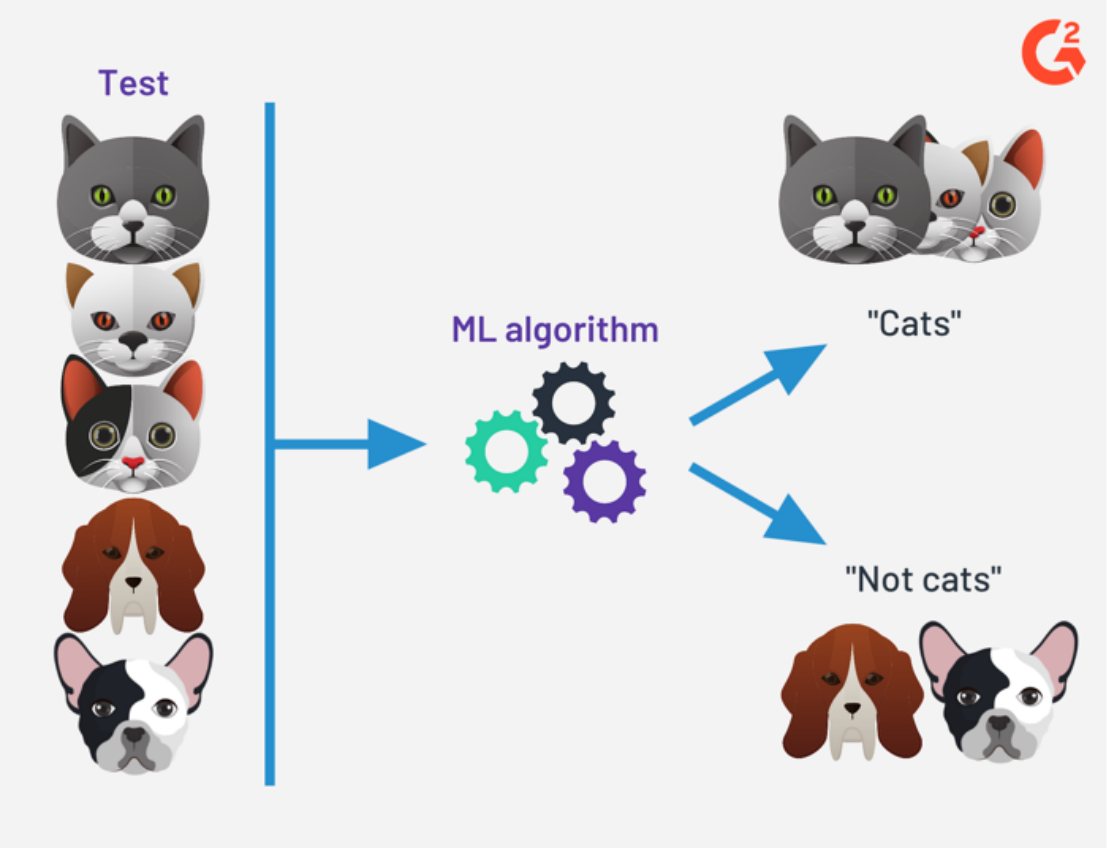
Aprendizado Não-Supervisionado

Aprendizado Supervisionado $\times$ Não-Supervisionado

- Supervisionado: rótulo de classe conhecido para cada item observado
- Não-supervisionado: rótulo de classe desconhecido; nenhuma informação sobre quantidade de classes

Aprendizado Supervisionado × Não-Supervisionado

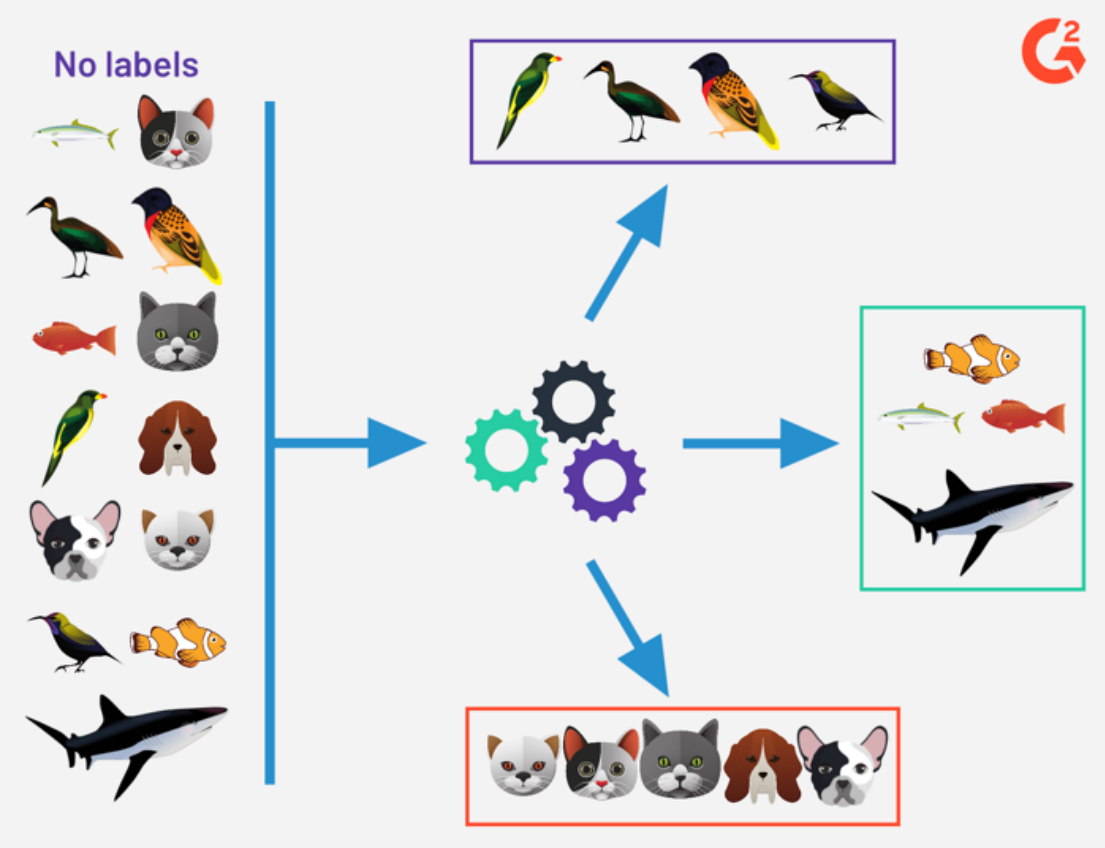
- Predição usando um classificador já treinado:



Fonte: https://learn.g2.com/supervised-vs-unsupervised-learning

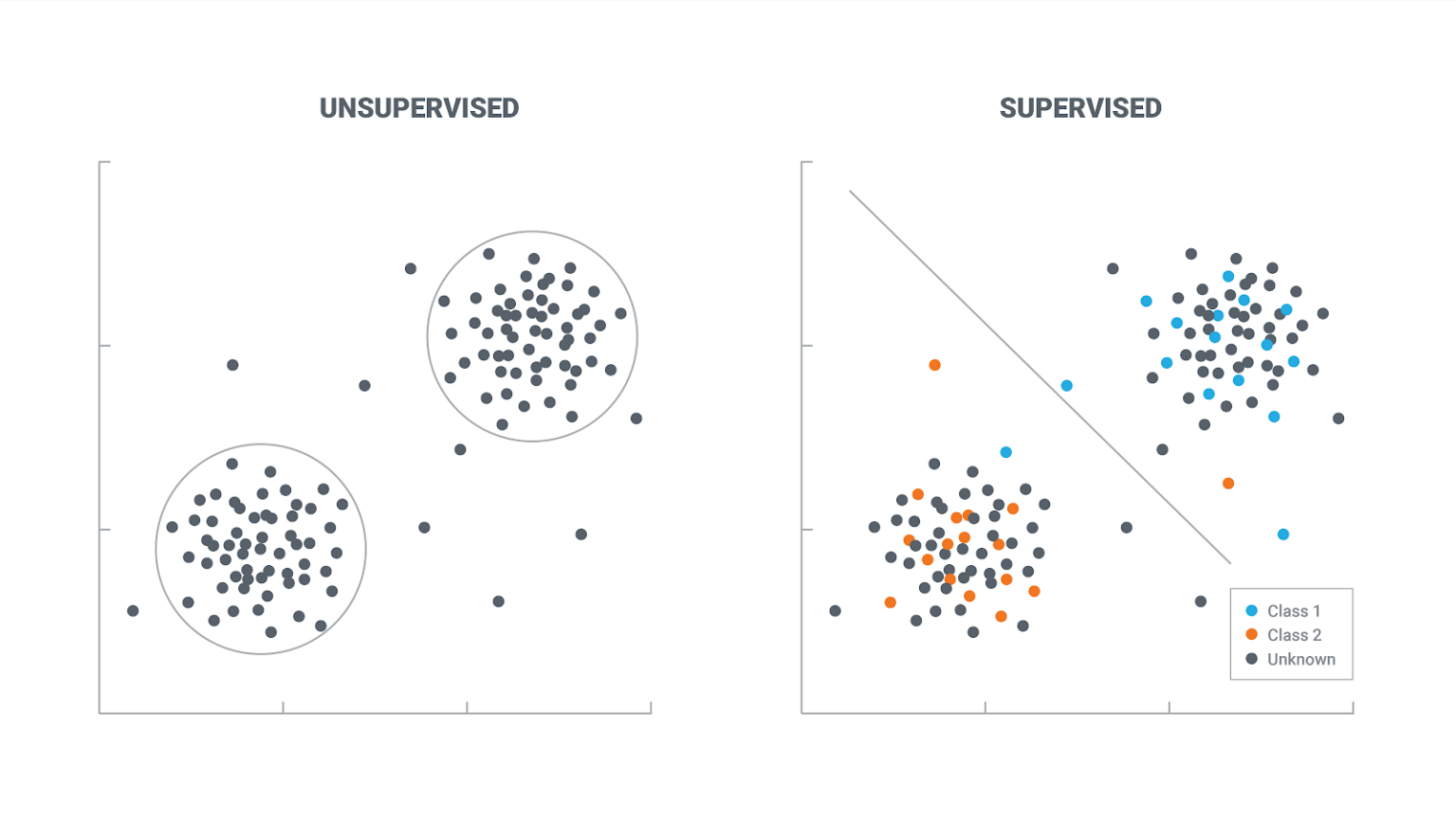
Aprendizado Supervisionado × Não-Supervisionado

- Clustering:



Fonte: https://learn.g2.com/supervised-vs-unsupervised-learning

Aprendizado Supervisionado × Não-Supervisionado



Fonte: https://lawtomated.com/supervised-vs-unsupervised-learning-which-is-better/

Clustering

- Clustering ou cálculo de aglomerados/agrupamentos é a principal abordagem usada para classificação não-supervisionada.
- Nenhuma informação sobre classes é disponível (no máximo, número de classes)
- Objetivo: encontrar grupos ou estruturas naturais "escondidas" no conjunto de dados.
- O conceito de grupo é vagamente definido; em geral envolve conceitos como similaridade (itens similares, itens parecidos)
- Um bom agrupamento é aquele no qual itens de um mesmo grupo são "mais similares" entre si e ao mesmo tempo "menos similares" aos itens de outros grupos.

Clustering

- O princípio básico do clustering é agrupar os itens baseado em alguma noção de similaridade
- Essas definições dependem da natureza e tipo dos dados (nominal, numérico, binário, contínuo, série temporal, cadeia, grafo, etc)
- Em geral podemos pensar em similaridade como distância entre pontos

Clustering - Métricas

- Tipos de medidas:
  - Entre itens: $d(\mathbf{x}_i, \mathbf{x}_j)$ (proximidade entre os pontos $\mathbf{x}_i$ e $\mathbf{x}_j$)
  - Entre itens e conjunto de itens: $d(\mathbf{x}, C)$ (proximidade entre um ponto $\mathbf{x}$ e um cluster C)
  - Entre dois conjuntos de itens: $d(C_i, C_j)$ (proximidade entre os clusters C_i e C_j)
  - Distância entre um item $\mathbf{x}$ e um cluster C:
    - Menor distância: $d(\mathbf{x}, C) = \min_{\mathbf{y} \in C} d(\mathbf{x}, \mathbf{y})$
    - Maior distância: $d(\mathbf{x}, C) = \max_{\mathbf{y} \in C} d(\mathbf{x}, \mathbf{y})$
    - Distância média: $d(\mathbf{x}, C) = \frac{1}{|C|} \sum_{\mathbf{y} \in C} d(\mathbf{x}, \mathbf{y})$

Clustering - Métricas

- Escolher um representante para o grupo e calcular a distância ponto-a-ponto:
  - **ponto** (faz sentido em grupos esféricos)
    - **ponto médio**: $\mathbf{m}_p = \frac{1}{|C|} \sum_{\mathbf{y} \in C} \mathbf{y}$
    - **ponto central**: $\mathbf{m}_c \in C$ tal que $\sum_{\mathbf{y} \in C} d(\mathbf{y}, \mathbf{m}_c) \leq \sum_{\mathbf{y} \in C} d(\mathbf{y}, \mathbf{m}), \forall \mathbf{m} \in C$
    - **ponto mediano** : $\mathbf{m}_{med}$ tal que $med\{d(\mathbf{y}, \mathbf{m}_{med}), \mathbf{y} \in C\} \leq med\{d(\mathbf{y}, \mathbf{m}), \mathbf{y} \in C\}, \forall \mathbf{m} \in C$
  - **hiperplano, hipercírculo** : a distância entre um item $\mathbf{x}$ e um cluster C, quando o **cluster é representado por uma hipercurva** pode ser simplesmente a distância do ponto à curva

Clustering - Abordagens

- Duas abordagens comuns:
  - Hierárquico: aglomerativo ou divisivo
  - Iterativo: ajuste iterativo de uma partição inicial, visando a minimização de algum custo

Clustering - Iterativo

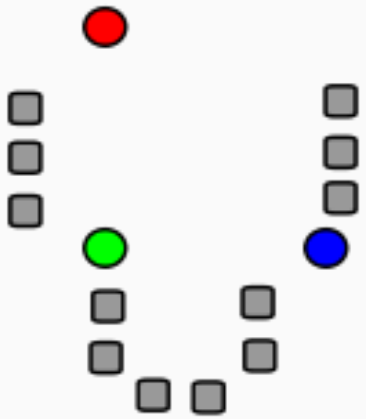
- Ideia geral:
  - Inicializar o processo com uma partição qualquer (por exemplo, escolhida aleatoriamente)
  - Repetir até algum critério de parada ser satisfeito
    - Modificar ligeiramente a partição atual
    - Verificar se a nova partição tem custo menor (segundo a função critério). Se tiver, substituir a solução por essa.
  - Devolver a melhor partição que foi obtida

k-Means Clustering

- k-Means é um dos algoritmos mais conhecidos:
  1. Escolher k pontos no espaço onde estão os itens a serem agrupados. Esses pontos corresponderão aos centróides iniciais dos grupos.
  2. Associar cada item ao grupo cujo centróide está mais próximo.
  3. Após todos os itens terem sido associados a algum grupo, recalcular os centróides de cada grupo
  4. Repetir os passos 2 e 3 até os centróides não mudarem mais de posição, ou até algum outro critério de parada ser atingido.

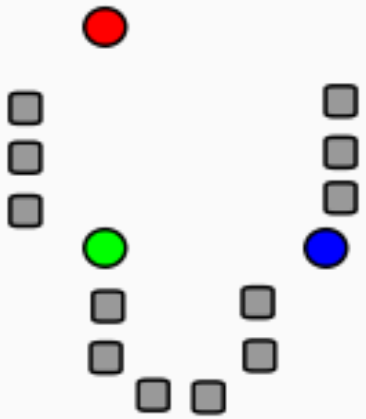
k-Means Clustering - Treinamento

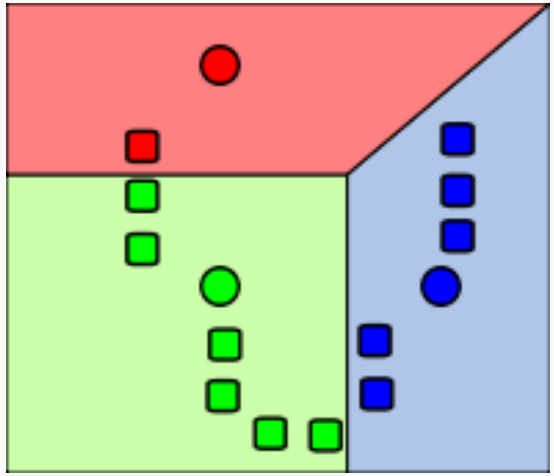
- Simulação ($k = 3$):



k-Means Clustering - Treinamento

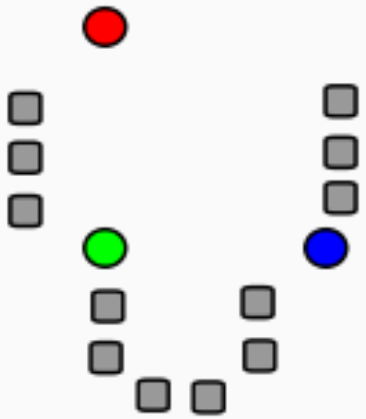
- Simulação ($k = 3$):

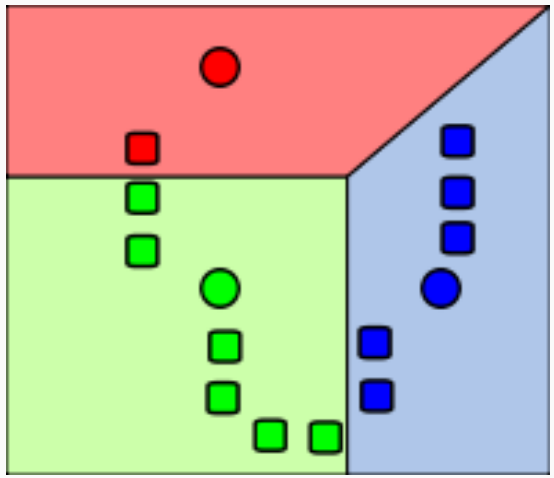


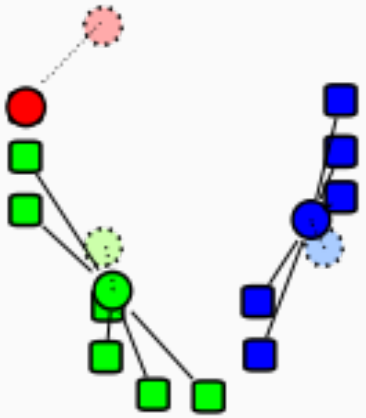


k-Means Clustering - Treinamento

- Simulação ($k = 3$):

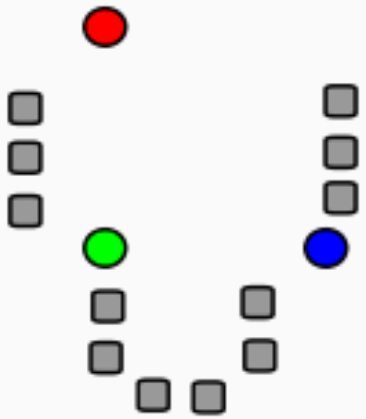


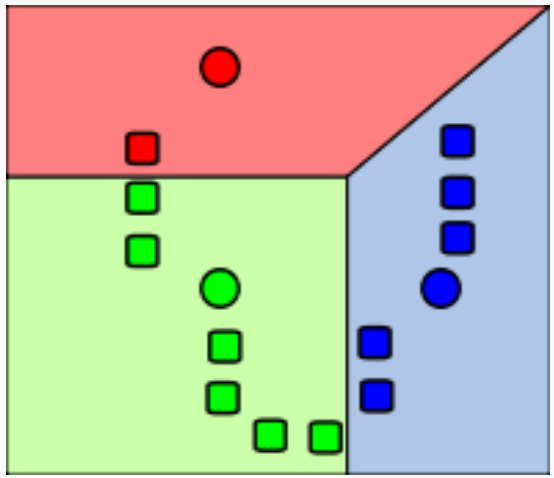


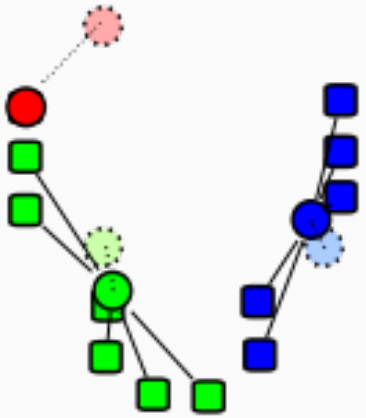


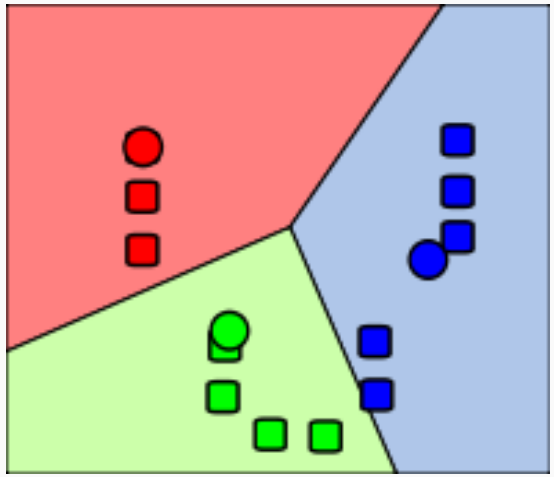
k-Means Clustering - Treinamento

- Simulação ($k = 3$):

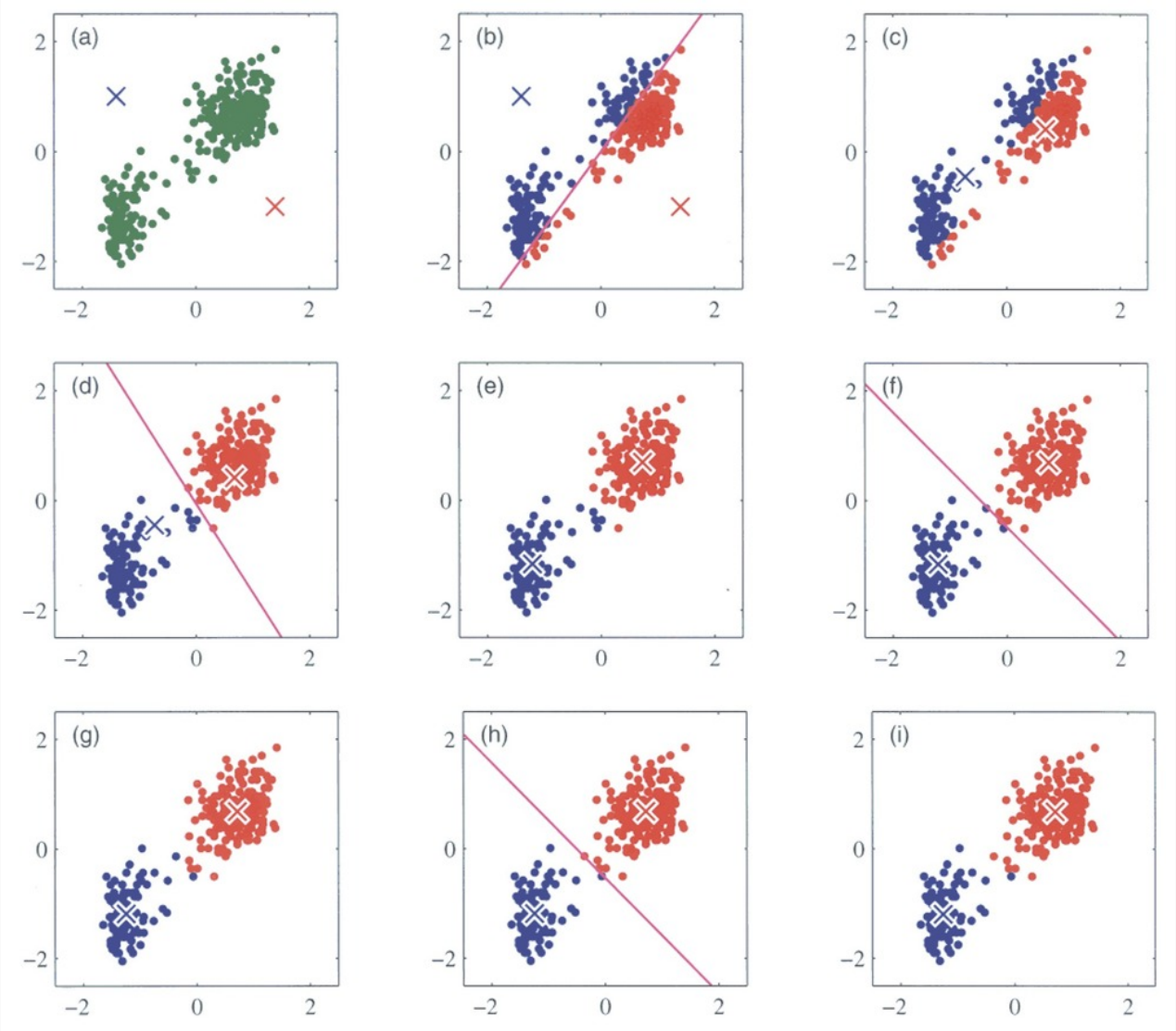






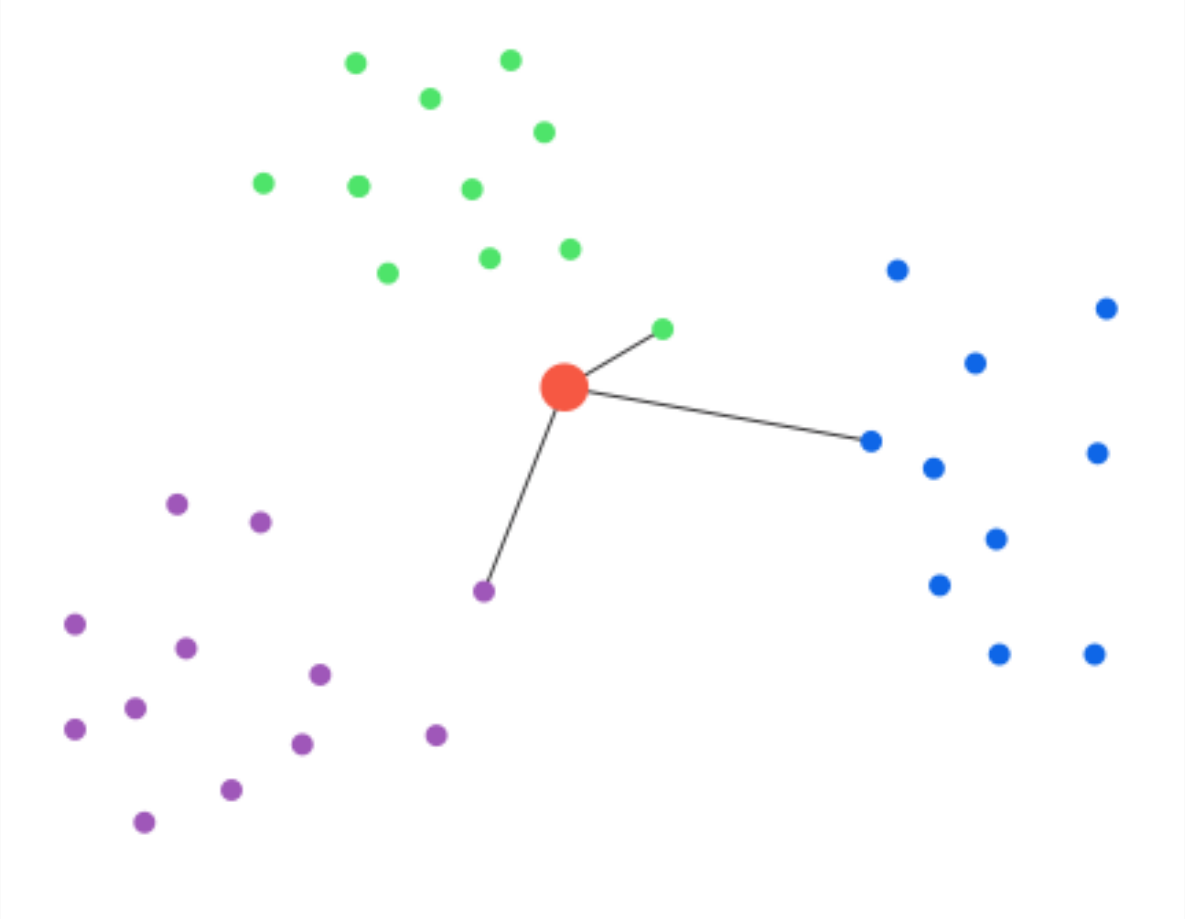


k-Means Clustering - Treinamento



Fonte: http://dendroid.sk/2011/05/09/k-means-clustering/

k-Means Clustering - Inferência/Predição



Fonte: http://dendroid.sk/2011/05/09/k-means-clustering/

Fim!