**Tecnicas de Agrupamento - notas da aula**

1. **Dois grandes grupos de tecnicas de ML:**

* aprendizado supervisionado
  + supervisionado pelo especialista do domínio (n o de dados)
  + a supervisão é por meio de rotulação de bases
    - O especialista é quem faz as inferências e te indica o que é para mandar para a base
    - Ao longo do tempo, o especialista monta um banco de dados suficiente para a execução de um modelo de ML
    - Pode ser uma pessoa, sistema, qualquer agente confiável para o processo de classificação.
    - Não apenas rótulos de classificação, mas em processos de regressão o target é numérico (continuo).
    - Podem haver mais de uma coluna (ténicas de multiclasse, mais de um Target)
* Aprendizado não supervisionado
  + Você entrega os dados para o partições (não modelos) e tenta extrair padrões comportamentais dentro do dataframe
  + Não geram modelos, geram partições!
  + em agrupamento, temos duas principais tendências nao supervisionadas
    - Associação
      * perdi o crl da explicação
    - Agrupamento de Dados
      * ele não sabe exatamente qual é o critério utilizado para esses agrupamentos
      * os atributos entregues serão os balizadores para tais classificações.
      * O algoritmo atribui importância com base nos atributos alimentados. pode ser encarado como “Características”
  + O que acontece na pratica é:
    - Uma base não rotulada passa por um processo de agrupamento
    - registra essas partições geradas, estuda e manipula elas a fim de afunilar os resultados
    - Chega através de analises nas partições finais
    - E ao fim, treina um modeo com base nos registros gerados
      * GRANDE PROBLEMA DISSO - Uma tecnica de agrupamento espera ERROS, então a criticidade do processo é extremamente relevante.
      * Ex classico: Itau - analise de curriculo, eles precisavam abrir uma vaga e conseguir ocupar a vaga onde 2k curriculos enviados. Tentaram fazer o melhor agrupamento de dois grupo (relevantes e nao relevantes)
        + Pode haver um cara que tem um otimo curriculo mas possui um elemento presente que pode acabar classificando erroneamente como um candidato ruim.
    - Uma coisa que nao faz sentido nenhum é pegar uma base rotulada, remover o target, e tentar fazer um agrupamento
      * NAO DA CERTO, pois o algoritimo de agrupamento NAO está olhando aquele target em específico e te retornará o que voce imagina kkkk

1. **Agrupamento Hard vs Soft**

Agrupamento hard = cada elemento so pode ter um cluster

Agrupamento soft = cada elemento oode pertencer a mais de um cluster.

1. Kmeans

Como o objetivo das técnicas é simular o mundo real, existem cenários onde o referencial é algo mutável. A altura de alguem, o que define se alguém é alto ou baixo está relacionado a o contexto, 1.80m é alto? Para altura de professores de uma instituição sim, mas para um time de basquete não.

Esse método analisa os pontos apresentados, definindo centroides aleatórios com base no numero de partições desejadas.

A partir do chute inicial ele faz um primeiro agrupamento com base no que está perto do centroide, reavalia e calcula um novo centroide.

As principais condições de parada do K Means são

* Ponto de convergencia (os resultados se aproximam bastante)
* Numero maximo de iterações

—-------------------

Agrupamento hieraquico

O objetivo do método é gerar um Dendograma, no notebook tem bons exemplos visuais.

Monta-se uma matriz de distancia com a distancia entre dois pontos