# Validação de modelos de clusterização

Autor(A): Patrícia Ferreira da Silva
Tutor: Luiz Fernando de Frias

# Validação de modelos de clusterização

- 1. Escolha da base de dados
- 2. Justificativa
- - Base sobre preços de casas
- - Agrupar preços de casas por região

- JUSTIFICAR O NÚMERO DE CLUSTER ESCOLHIDO
- O ponto 4 indicado no gráfico é o ponto ideal para escolha do número de clusters pois é o ponto que está mais próximo de seus pontos vizinhos.

# Validação de modelos de clusterização

- Compare os dois resultados, aponte as semelhanças e diferenças e interprete
- K meas
- - Utilizei o NearestNeighbors para escolher o melhor eps, e verificar se a distribuição dos dados ficaria melhor, mas, os gráficos é aparentemente igual, o k means dividiu em 4 grupos.
- - DBSCAN- recuperou em apenas um grupo e alguns pontos aleatorios (outliers)

# Validação de modelos de clusterização

- 3. **Escolha mais duas medidas de valição para comparar com o indice de silhueta**
- Indice de silhueta- valida o desempenho do agrupamento com base na diferença par a par das distâncias entre e dentro do agrupamento. Além disso, o número de cluster ideal é determinado maximizando o valor desse índice.
- DBCV- desenvolvido para superar a limitação de indice de silhueta na medição de qualidade de clusters não convexos e modelos baseado em densidade.
- CVNN- Índice de validação de agrupamento baseado em vizinhos mais próximos, avalia a separação intercluster com base em objetos que carregam as informações geométricas de cada cluster.

# Validação de modelos de clusterização

- **4. A silhueta é um índice indicado para escolher o número de clusters para o algoritmo DBScan?**
- Resp--- Quando não há rótulos disponíveis, é comum escolher uma métrica objetiva, como o Silhouette Score , para avaliar e, em seguida, decidir sobre o resultado final do agrupamento. O Silhouette Score mede a coesão e a separação do cluster com um índice entre -1 a 1. A distância não é aplicável para uma técnica baseada em densidade.
- 
- Isso significa que o Silhouette Score e índices semelhantes são inadequados para medir técnicas baseadas em densidade como o DBScan.

# Validação de modelos de clusterização

- **Medidas de similaridade**
- **1. Um determinado problema, apresenta 10 séries temporais distintas. Gostaríamos de agrupá-las em 3 grupos, de acordo com um critério de similaridade, baseado no valor máximo de correlação cruzada entre elas. Descreva em tópicos todos os passos necessários.**
-
- pegar ambas  e começar a dar um lag nas series, temos uma unidade de tempo anda com ele, calcular a correlação de pearson entre elas, faz um shift e calcula a correlação de novo e de novo e de novo.....ela cria uma curva que mostra a correlação de person, onde iremos saber o ponto ideal da sincronia e onde esta adiantando ou atrasado.

# Validação de modelos de clusterização

- 2. Para o problema da questão anterior, indique qual algoritmo de clusterização você usaria. Justifique
- - Clsterização hierárquica. assim poderia atribuir todos os pontos de dados a um cluster própio.
- 
- 3. Indique um caso de uso para essa solução projetada.
- - clusterizar bases que tenha relação com o tempo, por ex: temperatura anual, tempo de compra e venda, cotação do dolar.... assim podendo fazer a previsão de valores futuros, podendo enxergar um padrão na série que nos permita saber qual será a tedência desses valores.

- **4. Sugira outra estratégia para medir a similaridade entre séries temporais. Descreva em tópicos os passos necessários.**
- - Algoritmo DTW(Dynamic time warping).É utilizada para encontrar o alinhamento não-linear ótimo entre duas sequências de valores numéricos.
- - o primeiro passo é realizar o cálculo da matriz de pesos de dimensão, a partir do resultado da matriz de pesos é calculada a matriz de custos acumulados D, o k-ésimo caminho de menos custo D gera uma similaridade entre o padrão e a subsequência de V com a distância associada k, onde ak é o ponto inicial e bk é o ponto final da subsequência k. Cada ponto mínimo na última linha da matriz de custo acumulado, produz um alinhamento, onde K é o número mínimo de pontos na última linha da matriz de custos acumulados D.
- - Algoritmo dtwclust, por que ele utiliza técnicas relacionadas à distância do dynamic, utiliza implementações de agruppamentos particional, hierárquico e pode ser facilmente estendida com medidas de distância personalizadas e definições de centróides.
-