# Validação de modelos de  clusterização

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

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

- Justificar o número de clusters escolhidos
- A principal diferença entre os gráficos acima está no tamanho da silhueta de cada gráfico, os gráfico apresenta tamanhos semelhantes entre si, onde com cluster = 2 possui maior discrepância entre si, ja o com cluster = 4 estão com tamanhos semelhantes entre si, o que surge ser a melhor escolha para a quantidade de clusters.

- Compare os dois resultados, aponte as semelhanças e diferenças e interprete
- K meas recuperou 4 grupos distintos já o dbscan não conseguiu recuperar
- K MEAS
- fácil de ser implementado e interpretado
- * é mais escalável mais eficiente
- *requer que o usuário diga inicialmente o n° de clusters
- *sensível a outliers

- DBScan
- * simples e fácil de ser implementado
- * não requer que o usuário diga p n° de cluster
- *não é sensível a outlires
- 
- K meas é um algoritmo de aprendizado de máquina superimpressionados, ele identifica o número de k centróides e, em seguida
- aloca todos os pontos de dados para o cluster mais próximo, mantendo os centroides o menor possível.
- DVScan- é baseado em densidade de aplicativos com ruido, ele define os clusters como o maior conjunto de pontos densamente
- conectados, pode dividir regiões com densidade alta o suficiente em clusters e pode encontrar clusters de forma arbitrárias
- em banco de dados espaciais ruidosos.
- 

- Escolha mais duas medidas de valição para comparar com o indice de silhueta
- Dendograma
- KElbowVisualizer
- Comparação
- método da silhueta, é analisado um coeficiente resultante de um cálculo da distância entre os centróides, levando em consideração o agrupamento dos dados que os cerca, observando os graficos, o que melhor represnta é o gráfico com k= 4, estão com tamanhos de silhuenta mais semelhantes entre si

- O dendrograma é um diagrama de árvore que exibe os grupos formados por agrupamento de observações em cada passo e em seus níveis de similaridade, observando o gráfico ele agrupa os dados em quantro grandes grupos.
- Metodo do cotovelo KElbowVisualizer, temos o valor de quão próximo eles são uns dos outros, observando o grafico, ele aponta paraum numero ideal de 3 ou 4 clustrs, sendo k =4 com melhor tempo de resposta para o algoritmo.

- Para o BDScan, a medida de validação melhor é o DBCV pois funciona para algoritmos de agrupamentos baseados em desidade precisamente porque leva em cosideração o ruído e captura a propriedade de forma dos agrupamentos por meio de densidadese não por distância.
- Analisando todos os resultados o melhor K para o olgoritmo seria K = 4

- 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 Silhueta Score , para avaliar e, em seguida, decidir sobre o resultado final do agrupamento. O SilSilhueta Score mede a coesão e a separação do cluster com um índice entre -1 a 1. Ele NÃO leva em consideração o ruído no cálculo do índice e faz uso de distâncias. A distância não é aplicável para uma técnica baseada em densidade. Não incluir um ruído no cálculo da métrica objetiva viola uma suposição inerente ao agrupamento baseado em densidade.
- 
- Isso significa que o Silhueta Score e índices semelhantes são inadequados para medir técnicas baseadas em densidade como o DBScan.

- # 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.

- 2. Para o problema da questão anterior, indique qual algoritmo de clusterização você usaria. Justifique
- KNN, ele é utilizado para fazer estimativas de densidade, permitindo verificar quais são as regiões de alta e baixa densidade, retornando um índice com valor de -1 a 1.
-
- Algoritmo dtwclust, por que ele utiliza técnicas relacionadas à distância do dynamic, utiliza implementações de agrupamentos particional, hierárquico e pode ser facilmente estendida com medidas de distância personalizadas e definições de centroides

- 3. Indique um caso de uso para essa solução projetada.
- Pode- se usar esse caso para clusterizar bases que tenha relação com o tempo, por ex: temperatura anual, tempo de compra e venda....
- 4. Sugira outra estratégia para medir a similaridade entre séries temporais. Descreva em tópicos os passos necessários.
- similaridade= agrupar o moviento dela de acordo com o indice temporal(movemento da temperatura ao logo do tempo, subir e descer) essa é a similaridade com os outros grupos (sincronia) utiliza correlação de pearson para calcular similaridadevalor maximo é entre -1 e 1.
-