Escolha de base de dados

Para as questões a seguir, usaremos uma base de dados e faremos a análise exploratória dos dados, antes da clusterização.

1. Escolha uma base de dados para realizar o trabalho. Essa base será usada em um problema de clusterização.
   https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering

2. Escreva a justificativa para a escolha de dados, dando sua motivação e objetivos.

   Motivação

   Tratar esses dados poderia entender mais sobre as características químicas dos vinhos e como elas afetam suas classificações, como consequência disso podemos usar essas informações para produzir vinhos com melhores classificações, qualidade e com as melhores características para os consumidores

   Objetivo

   Criar um modelo que preveja a classificação de um vinho com base em suas características químicas. Esse modelo será criado utilizando a técnica K-means para dividir os vinhos em grupos com base em suas características químicas e ver se esses grupos têm alguma relação com as classificações dos vinhos.

3. Mostre através de gráficos a faixa dinâmica das variáveis que serão usadas nas tarefas de clusterização. Analise os resultados mostrados. O que deve ser feito com os dados antes da etapa de clusterização?

   Gráficos apresentados no notebook jupyter.

   <u>Análise dos Resultados:</u>

   Gráfico Pairplot
   Podemos observar que muitas dessas distribuições fogem da normalidade e fica claro a presença de outliers por visualizar varios pontos espalhados. Por isso, para que o modelo tenha melhor desempenho, é necessária a padronização da escala das variáveis para que seus valores fiquem mais próximos a média e se aproximem a uma distribuição normal.

   Gráfico Correlação
   A correlação alta entre Total_Phenols x Flavonoids (0.86) faz todo sentido, uma vez que os flavonóides também são compostos que fazem parte da classe dos fenóis. Assim sendo, muito provavelmente as duas variáveis explicam a mesma coisa.

   OD280 x Total_Phenols (0,70) e OD280 x Flavonoids (0,79). As relações podem indicar que essas variáveis explicam aspectos parecidos do vinho ou o mesmo aspecto.

   O que deve ser feito com os dados antes da etapa de clusterização?

Verificar se os dados estão completos e limpos: é importante verificar se os dados não contêm valores faltantes ou incorretos, pois isso pode afetar negativamente o resultado da clusterização.
Normalizar ou padronizar as variáveis: se as variáveis tiverem escala diferente, isso pode afetar o resultado da clusterização. Portanto, é importante normalizar ou padronizar as variáveis para que elas estejam na mesma escala.
Escolher o número de clusters: antes de realizar a clusterização, é importante decidir qual o número de clusters a ser utilizado

4. Realize o pré-processamento adequado dos dados. Descreva os passos necessários.

   Resultados apresentados no jupyter notebook.

   - Verifica se tem dados nulos
   - Verifica se tem dados duplicados
   - Normalização dos dados
   - Transformação de variáveis: em alguns casos, pode ser necessário aplicar transformações em uma ou mais variáveis para torná-las mais adequadas para a análise

Clusterização

Para os dados pré-processados da etapa anterior você irá:

1. Realizar o agrupamento dos dados, escolhendo o número ótimo de clusters. Para tal, use o índice de silhueta e as técnicas:
    1. K-Médias
    2. DBScan

    Com os resultados em mão, descreva o processo de mensuração do índice de silhueta. Mostre o gráfico e justifique o número de clusters escolhidos.

Gráficos apresentados no notebook jupyter

Processo de mensuração do índice de silhueta:

O índice de silhueta varia de -1 a 1, onde valores próximos de -1 indicam que o ponto de dados é muito melhor atribuído ao outro cluster, enquanto valores próximos de 1 indicam que o ponto de dados é muito melhor atribuído ao seu próprio cluster. Valores próximos de 0 indicam que o ponto de dados não está muito claramente atribuído a um cluster em particular.

Para calcular o índice de silhueta de cada ponto de dados, siga estes passos:

Calcule a distância entre cada ponto de dados e todos os outros pontos de dados do mesmo cluster. A média dessas distâncias é chamada de "coesão" do cluster.

Calcule a distância entre cada ponto de dados e todos os pontos de dados do cluster mais próximo. A média dessas distâncias é chamada de "separação" do cluster.

Calcule o índice de silhueta para cada ponto de dados usando a seguinte fórmula:

índice de silhueta = (separação do cluster - coesão do cluster) / máximo(separação do cluster, coesão do cluster)

Calcule a média dos índices de silhueta para todos os pontos de dados para obter o índice de silhueta médio para o conjunto de dados.

Número de clusters escolhidos: O valor de K=3 é o mais adequado pois é ponto mais alto da silhueta

2. Compare os dois resultados, aponte as semelhanças e diferenças e interprete.

   Resultados apresentados no notebook jupyter

   No k-Means com a escolha de 3 clusters com o índice de silhueta, podemos observar que tivemos os clusters vermelho e verde acima da média global e bem distribuído, já o cluster azul possui ruídos e quase abaixo da média global. O k-means agrupado com 2 cluster notamos uma melhora na distribuição e a ausência de ruídos. No DBScan que é um modelo baseado em densidade observamos um cluster com uma grande quantidade de ruídos.

observa-se que nele existe cluster negativo, isso acontece devido a presença de outliers no dataste, desta forma, é necessário a retirada desses outliers para melhorar ainda mais a clusterização através do método K-Means.

Por fim, observa-se que no modelo k-Means existe cluster negativo com a escolhe de 3 clusters, isso acontece devido a presença de outliers no dataste, desta forma, é necessário a retirada desses outliers para melhorar ainda mais a clusterização através do método K-Means.

3. Escolha mais duas medidas de validação para comparar com o índice de silhueta e analise os resultados encontrados. Observe, para a escolha, medidas adequadas aos algoritmos.

   Resultados apresentados no notebook jupyter

4. Realizando a análise, responda: A silhueta é um o índice indicado para escolher o número de clusters para o algoritmo de DBScan?

   Não, a silhueta não é o índice mais adequado para escolher o número de clusters para o algoritmo DBScan. O DBScan é um algoritmo de agrupamento baseado em densidade, o que significa que ele encontra clusters de alta densidade em um conjunto de dados. Ele não requer que você especifique o número de clusters antes de começar o processo de agrupamento, pois ele pode encontrar clusters de qualquer tamanho e forma. O índice de silhueta é um índice que pode ser usado para avaliar a qualidade de um agrupamento em um conjunto de dados, mas ele não é um parâmetro do algoritmo 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.

   1º Passo: Colete as séries temporais para o problema em questão. Certifique-se de ter suficientes observações para cada série temporal para que os cálculos de correlação sejam precisos.

   2º Passo: Calcule a correlação cruzada entre todas as combinações de séries temporais. A correlação cruzada mede a similaridade entre duas séries temporais ao longo do tempo.

   3º Passo: Ordene as séries temporais de acordo com o valor máximo de correlação cruzada entre elas. As séries temporais com os valores mais altos de correlação cruzada são mais similares entre si do que aquelas com valores mais baixos.

   4º Passo: Agrupe as séries temporais em 3 grupos. Você pode usar um método de agrupamento hierárquico ou um método de agrupamento baseado em centroides para fazer isso. O método de agrupamento hierárquico cria um diagrama de árvore que mostra como as séries temporais são relacionadas entre si, enquanto o método de agrupamento baseado em centroides cria grupos com base em características comuns entre as séries temporais.

   5º Passo: Avalie o resultado do agrupamento. Verifique se as séries temporais em cada grupo são realmente similares entre si e se o

agrupamento reflete o critério de similaridade que você está tentando medir. Se necessário, ajuste o método de agrupamento ou o critério de similaridade para obter um resultado mais satisfatório.

2. Para o problema da questão anterior, indique qual algoritmo de clusterização você usaria. Justifique.

   Para agrupar as séries temporais em 3 grupos de acordo com um critério de similaridade baseado no valor máximo de correlação cruzada entre elas, eu recomendaria usar o método de agrupamento hierárquico.

   O método de agrupamento hierárquico é um algoritmo de clusterização que cria um diagrama de árvore, chamado de dendrograma, que mostra como as séries temporais são relacionadas entre si. Ele começa agrupando as séries temporais mais similares entre si e, a partir daí, vai combinando os grupos formados até que todas as séries temporais estejam agrupadas em um único grupo.

   O método de agrupamento hierárquico é uma opção boa para o problema em questão, pois permite visualizar de forma clara como as séries temporais estão relacionadas entre si e como elas foram agrupadas. Além disso, o dendrograma gerado pelo método de agrupamento hierárquico pode ser facilmente cortado em qualquer ponto para formar grupos de tamanho desejado, o que é útil para criar os 3 grupos desejados.

3. Indique um caso de uso para essa solução projetada.

   Uma possível aplicação para a solução projetada seria na análise de séries temporais de vendas de produtos em uma empresa. Suponha que a

empresa tenha 10 produtos diferentes e queira saber quais produtos têm comportamentos de vendas semelhantes entre si. A solução projetada poderia ser usada para agrupar esses produtos em 3 grupos de acordo com o critério de similaridade baseado no valor máximo de correlação cruzada entre as séries temporais de vendas de cada produto. Isso permitiria à empresa identificar padrões comuns entre os produtos e tomar decisões de gerenciamento de produtos mais informadas.

4. Sugira outra estratégia para medir a similaridade entre séries temporais. Descreva em tópicos os passos necessários.

   Uma estratégia alternativa para medir a similaridade entre séries temporais seria usar a distância de Euclides. Aqui estão os passos gerais que você pode seguir para medir a similaridade entre séries temporais usando a distância de Euclides:

   1º Passo: Colete as séries temporais para o problema em questão. Certifique-se de ter suficientes observações para cada série temporal para que os cálculos de distância sejam precisos.

   2º Passo: Calcule a distância de Euclides entre todas as combinações de séries temporais. A distância de Euclides é uma medida de similaridade entre duas séries temporais que leva em conta a distância entre os pontos em um espaço multidimensional.

   3º Passo: Ordene as séries temporais de acordo com a distância de Euclides entre elas. As séries temporais com as menores distâncias de Euclides são mais similares entre si do que aquelas com distâncias maiores.

4º Passo: Escolha um limiar de similaridade. O limiar de similaridade é o valor mínimo de distância de Euclides que as séries temporais devem ter para serem consideradas similares entre si. Você pode escolher um limiar de similaridade de acordo com suas necessidades e objetivos específicos.

5º Passo: Agrupe as séries temporais que atendem ao limiar de similaridade. As séries temporais que têm distâncias de Euclides abaixo do limiar de similaridade são agrupadas juntas, enquanto as séries temporais com distâncias de Euclides acima do limiar são deixadas em grupos separados.

6º Passo: Avalie o resultado do agrupamento. Verifique se as séries temporais em cada grupo são realmente similares entre si e se o agrupamento reflete o critério de similaridade que você está tentando medir. Se necessário, ajuste o limiar de similaridade ou o método de agrupamento para obter um resultado mais satisfatório.