escola britânica de artes criativas & tecnologia

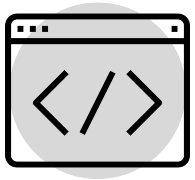
Profissão Cientista de Dados M29

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

K-means

- Explore o Primeiro Contato com a Técnica
- Compreenda a Distância
- Entenda como Funciona o Algoritmo
- Realize a Aplicação
- Determine o Número de Clusters



Explore o Primeiro Contato com a Técnica

- Ao trabalhar com técnicas de agrupamento como o K-means, é importante selecionar apenas as variáveis numéricas para análise. Variáveis categóricas podem não ser adequadas para esse tipo de técnica.
- Realize uma análise descritiva básica e visualize os dados através de gráficos antes de aplicar o algoritmo. Isso pode ajudar a identificar padrões visíveis nos dados e sugerir a presença de grupos distintos.

Explore o Primeiro Contato com a Técnica

- Ao executar o algoritmo K-means, defina o número de clusters e outros parâmetros com cuidado. Essas escolhas podem ter um grande impacto nos resultados do agrupamento.
- Lembre-se de que a definição de similaridade é crucial no algoritmo K-means. A próxima discussão deve ser sobre como definir similaridade e como melhorar os resultados do agrupamento.
- Após executar o algoritmo, inspecione os resultados e visualize os grupos em gráficos. Se os grupos não forem claramente distintos em todas as variáveis, pode ser necessário ajustar os parâmetros ou considerar outras técnicas de agrupamento.

Compreenda a Distância

- Sempre considere a escala das variáveis ao usar algoritmos que dependem de medidas de distância, como o K-means. Variáveis com escalas diferentes podem afetar os resultados de maneira indesejada.

- Ao apresentar os resultados de um algoritmo de agrupamento, é útil comparar os grupos resultantes com categorias conhecidas, se disponíveis. Isso pode ajudar a interpretar os grupos e verificar a qualidade do agrupamento.

Entenda Como Funciona o Algoritmo

- Ao calcular a distância de cada ponto a cada centróide, é importante classificar cada ponto ao grupo cujo centróide é o mais próximo. Isso é fundamental para o funcionamento do algoritmo K-means.
- Embora o algoritmo K-means possa identificar grupos de dados, a avaliação da qualidade desses grupos deve estar alinhada com o objetivo de negócio. Portanto, é importante ter uma compreensão clara do objetivo de negócio antes de avaliar os resultados do algoritmo K-means.
- Os centróides devem ser redefinidos de acordo com a média de cada uma das coordenadas dos pontos em seus respectivos grupos. Este processo deve ser repetido até que os centróides não se movam mais, indicando que o algoritmo convergiu.

Realize a Aplicação

- Entenda o negócio: Antes de realizar qualquer análise, é crucial entender o negócio e o contexto no qual você está trabalhando. Isso permitirá que você faça perguntas relevantes e obtenha insights úteis.
- Planeje com antecedência: Tenha em mente que a escolha do número ideal de grupos é uma decisão importante que deve ser tomada com cuidado. Planeje discutir e explorar essa questão em detalhes.

Determine o Número de Clusters

- Não busque cegamente o número "ótimo" de clusters: Embora existam métodos que ajudam a determinar o número ideal de clusters, lembre-se de que diferentes métricas podem resultar em diferentes números de clusters ideais. Portanto, é importante considerar vários indicadores e realizar uma análise estatística descritiva dos grupos para determinar se eles fazem sentido e diferem significativamente uns dos outros.

- Considere o objetivo do projeto: O número de clusters desejados pode ser influenciado pelo objetivo do projeto. Portanto, sempre tenha em mente o que você espera alcançar com a análise de agrupamento.

Determine o Número de Clusters

- Utilize o método do cotovelo e o método da silhueta: Esses dois métodos podem ajudar a determinar o número de clusters. O método do cotovelo envolve a criação de um gráfico da soma dos quadrados das distâncias contra o número de clusters, enquanto o método da silhueta calcula uma pontuação para cada ponto que mede quão semelhante é a outros pontos em seu cluster em comparação com pontos em outros clusters.

- Visualize e interprete os grupos resultantes: Use gráficos de barras, gráficos de médias padronizadas e gráficos de componentes principais para visualizar e interpretar os grupos. A escolha final do número de clusters deve ser baseada em uma combinação de métricas estatísticas, visualizações e relevância para o objetivo do projeto.

escola britânica de artes criativas & tecnologia

Bons estudos!