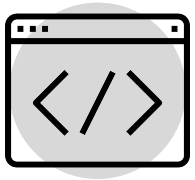
escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

Hierárqicos/ aglomerados



- Conheça o Scikit-learn
- Crie algoritmo
- Domine variáveis qualitativas
- Diferencie qualidade de complexidade
- Selecione modelos
- Regularização

Conheça o Scikit-learn

- Ao trabalhar com métodos de agrupamento, é importante selecionar apenas as variáveis numéricas e excluir os valores ausentes. Isso garantirá que o algoritmo funcione corretamente.
- Sempre visualize os resultados do seu agrupamento. Isso pode ajudá-lo a entender melhor como o algoritmo está agrupando seus dados e se esses grupos fazem sentido com base em seu conhecimento do conjunto de dados.
- Ao usar o método de agrupamento aglomerativo, lembre-se de que você precisa escolher o tipo de ligação e o limite de distância para o algoritmo. Essas escolhas podem ter um grande impacto nos resultados do seu agrupamento.
- Depois de ajustar o modelo, adicione os números do cluster ao conjunto de dados original. Isso permitirá que você visualize os resultados do agrupamento e compare-os com quaisquer categorias existentes em seus dados.

Crie algoritmo

- Ao agrupar elementos, é necessário redefinir a distância de todos os outros pontos a esse grupo. Existem vários critérios para redefinir essa distância, portanto, escolha o critério que melhor se adapta ao seu conjunto de dados e ao problema que você está tentando resolver.

- O algoritmo de agrupamento hierárquico é iterativo, o que significa que ele repete os passos de encontrar a menor distância e agrupar os pontos correspondentes até que todos os pontos estejam agrupados em um grande grupo. Esteja ciente de que isso pode ser demorado para grandes conjuntos de dados.

Crie algoritmo

- O algoritmo de agrupamento hierárquico é útil para entender a estrutura dos dados e pode ser uma ferramenta poderosa para a análise de dados. No entanto, como qualquer algoritmo, ele tem suas limitações e não é adequado para todos os problemas.

- A implementação do algoritmo em Python pode ser feita com pacotes como 'distance' do módulo 'spal'. Portanto, familiarize-se com essas bibliotecas e suas funções para implementar efetivamente o algoritmo.

Visualize pelo Dendograma

- Ao criar um dendrograma, certifique-se de entender o que os eixos X e Y representam. O eixo X geralmente representa o índice ou rótulo de cada ponto, enquanto o eixo Y representa a distância de agrupamento.

- Lembre-se de que a interpretação de um dendrograma é subjetiva e pode variar dependendo do objetivo da análise. Portanto, é importante ter clareza sobre o objetivo da sua análise antes de interpretar um dendrograma.

- Ao interpretar um dendrograma, use a distância no eixo Y para ajudar a determinar o número de grupos em seus dados. Se a distância para quebrar de um número de grupos para outro é pequena, pode ser útil considerar a divisão em mais grupos. No entanto, se a distância para quebrar de um número de grupos para outro é grande, pode ser melhor ficar com menos grupos.

Conheça os tipos de ligação

- Compreenda os diferentes tipos de ligação: Como cientista de dados, é crucial entender os diferentes tipos de ligação - ligação simples, ligação completa, ligação média e ligação de Ward - e como eles definem a distância de um ponto a um agrupamento. Cada tipo tem suas próprias características e pode levar a diferentes resultados de agrupamento.

- Escolha o tipo de ligação apropriado para o seu conjunto de dados: Não existe um "melhor" tipo de ligação. A escolha do tipo de ligação depende do tipo de agrupamento que você deseja obter. Portanto, é importante considerar o contexto e o objetivo do seu projeto ao escolher o tipo de ligação.

Conheça os tipos de ligação

- A interpretação dos resultados do agrupamento é uma habilidade importante para um cientista de dados. Pratique a interpretação dos resultados de diferentes tipos de ligação em diferentes conjuntos de dados para melhorar essa habilidade.

- Considere a ligação de Ward para uma abordagem mais estatística: Se você está procurando uma abordagem mais estatística para o agrupamento, considere usar a ligação de Ward, que é baseada no conceito da soma de quadrados das distâncias.

Aplique a distância

- A distância euclidiana, que é calculada usando o teorema de Pitágoras, é uma medida de distância comum, mas pode não ser a melhor escolha para todos os conjuntos de dados.
- Sempre teste diferentes medidas de distância para ver qual delas fornece os melhores resultados para o seu conjunto de dados específico.
- A distância de Manhattan, também conhecida como City block, é uma alternativa útil quando os deslocamentos são apenas horizontais e verticais. Independentemente do caminho escolhido, a distância total percorrida será a mesma.

Agrupe com dados mistos

- Trate os dados mistos de maneira semelhante ao tratamento de dados ausentes. Você pode optar por descartar a linha ou coluna ou preencher com alguma regra.
- Para variáveis qualitativas, crie variáveis dummy. Isso permitirá que o algoritmo trate essas variáveis de maneira adequada.
- Crie uma lista para informar o algoritmo sobre as variáveis qualitativas. Isso é importante porque o algoritmo trata essas variáveis de maneira diferente.
- Calcule a matriz de distância e execute o agrupamento. Lembre-se de que os métodos hierárquicos funcionam alimentando-os apenas com a matriz de distância, eliminando a necessidade da base de dados original para executar o algoritmo. No entanto, a base original ainda é necessária para avaliar o agrupamento.

Agrupe com dados mistos

- Lembre-se de que o objetivo do método de agrupamento não é prever uma variável específica, mas encontrar um padrão nos dados.
- O processo de agrupamento com dados mistos é escalável para bases de dados maiores. Portanto, não hesite em aplicá-lo em conjuntos de dados de grande escala.

Classifique

- Ao classificar novos pontos em grupos, calcule a distância a cada centroide. A menor distância indica o grupo ao qual o indivíduo pertence.
- Para métodos hierárquicos que não possuem um comando 'predict' natural, utilize um método de classificação, como o Random Forest. Este método é fácil de usar, robusto e eficaz.
- Ao preparar a base de dados para rodar o Random Forest, separe as variáveis explicativas (x) e a variável resposta (o grupo).
- Ao rodar o Random Forest e interpretar os resultados, lembre-se que é esperado um bom resultado, pois os grupos foram definidos exatamente com as variáveis usadas no Random Forest.

Classifique

- Ajuste a árvore com o melhor parâmetro na base de treinamento inteira e calcule a acurácia na base de teste.
- Esteja ciente de que mudanças na base de dados devido à variabilidade e mudanças no mundo podem afetar o agrupamento.
- Para métodos hierárquicos, é necessário um algoritmo de classificação.

escola britânica de artes criativas & tecnologia

Bons estudos!