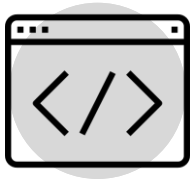
escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

Árvores I



- Entenda a árvore de decisão
- Carregue e trate dados
- Entenda a quebra de variável contínua
- Explore overfitting
- Domine overfitting, treino e teste
- Explore a poda da árvore
- Acompanhe os pós e contras

Entenda a árvore de decisão

- **Saiba quando usar classificação e regressão**: A classificação é usada quando o resultado é binário, como bom ou mau, enquanto a regressão é usada quando o resultado é contínuo, como prever o preço de um item.

Carregue e trate dados

- Sempre carregue e verifique o conjunto de dados antes de começar a trabalhar com ele. Isso inclui entender o que cada variável representa.
- Identifique e trate linhas duplicadas. Linhas duplicadas podem distorcer a análise e levar a conclusões errôneas.
- Verifique se há dados ausentes. Dados ausentes podem causar erros em muitos algoritmos de aprendizado de máquina, incluindo árvores de decisão.
- Reindexe o conjunto de dados após a remoção de linhas. Isso garante que os índices correspondam ao número de linhas.

Carregue e trate dados

- Trate os dados ausentes de maneira adequada. Dependendo do algoritmo que você está usando e da natureza dos dados, você pode preencher os dados ausentes com um valor específico, a média, a mediana, ou até mesmo excluir as linhas ou colunas com dados ausentes.

- Transforme variáveis categóricas em variáveis numéricas quando necessário. Alguns algoritmos de aprendizado de máquina não suportam variáveis categóricas.

- Divida o conjunto de dados em variáveis explicativas e a variável alvo. Isso é essencial para a modelagem de aprendizado de máquina.

Entenda a quebra de variável contínua

- Ao calcular o coeficiente de Gini, comece ordenando os dados e calculando a média dos pontos. Em seguida, determine o ponto de quebra e calcule as folhas a partir desse ponto.
- Compare os coeficientes de Gini para diferentes pontos de corte para determinar o melhor ponto de corte para o conjunto de dados.
- Utilize o conceito de média ponderada ao calcular o coeficiente de Gini para um ponto de corte. Calcule a média ponderada usando o número de amostras em cada folha e o coeficiente de Gini para cada folha.
- Lembre-se de que o ponto de corte determinado será usado ao construir a árvore de decisão para o conjunto de dados. Portanto, é crucial escolher o ponto de corte mais apropriado.

Explore overfitting

- Para evitar overfitting, é recomendável dividir o conjunto de dados em um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento é usado para treinar o modelo, enquanto o conjunto de teste é usado para avaliar a performance do modelo.
- A matriz de confusão é uma ferramenta útil para avaliar a performance do modelo. Ela mostra a quantidade de acertos e erros do modelo, permitindo identificar áreas onde o modelo pode ser melhorado.
- Se o modelo estiver sofrendo de overfitting, pode ser útil simplificar o modelo. Isso pode envolver a remoção de recursos desnecessários ou a redução da complexidade do modelo.

Domine overfitting, treino e teste

- Ao construir uma árvore de classificação, é importante considerar os parâmetros que serão passados durante a declaração. O critério, por exemplo, pode ser Gini ou entropia, e determina a qualidade da quebra.

- Ao calcular a acurácia para o conjunto de dados de treinamento, é importante ter uma sensibilidade para quando a acurácia é muito alta. Isso pode indicar overfitting, o que significa que o modelo pode não generalizar bem para novos dados.

- Para evitar o overfitting, é recomendado dividir o conjunto de dados em treinamento e teste. Isso pode ser feito usando a função train_test_split do Scikit-learn.

Domine overfitting, treino e teste

- Após a divisão dos dados, é importante treinar a árvore de decisão usando o conjunto de treinamento e prever os resultados usando o conjunto de teste. Isso permite avaliar como o modelo se comporta com dados que não foram usados durante o treinamento.
- É importante calcular a acurácia do modelo usando a matriz de confusão e a função accuracy_score do Scikit-learn. Isso fornece uma medida quantitativa de quão bem o modelo está realizando suas previsões.
- É crucial comparar a acurácia do modelo nos conjuntos de treinamento e teste. Se a acurácia for significativamente maior no conjunto de treinamento, isso pode ser um sinal de overfitting. Nesse caso, pode ser necessário ajustar o modelo ou usar técnicas de regularização para melhorar sua capacidade de generalização.

Explore a poda da árvore

- Considere limitar a profundidade máxima da árvore como uma forma de poda. Por exemplo, se a profundidade máxima for determinada como dois, a última linha de nós será podada e os nós anteriores se tornarão as folhas da árvore de decisão.
  Outra opção é limitar a quantidade máxima de amostras em cada folha. Por exemplo, se o número de amostras for limitado a três, qualquer folha com menos de três amostras será podada, resultando em uma nova configuração da árvore de decisão.

Acompanhe os pós e contras

- As árvores de decisão podem ser instáveis, portanto, é importante validar o modelo com diferentes conjuntos de dados para garantir que ele seja robusto.
- As árvores de decisão podem ser enviesadas se uma classe dominar o conjunto de dados. Portanto, é importante garantir que os dados estejam bem balanceados antes de treinar o modelo.

escola britânica de artes criativas & tecnologia

Bons estudos!