escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

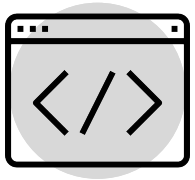
escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

escola britânica de artes criativas & tecnologia

Árvores II

- Compreenda a classificação multinominal
- Construa a classificação multinominal
- Conheça o cross-validation
- Analise os tipos de cross-validation
- Conheça o K-fold



Compreenda a classificação multinominal

- Use técnicas sofisticadas apenas quando necessário: Não aplique técnicas complexas apenas por sua sofisticação. Use-as quando forem necessárias para resolver um problema específico.

- A métrica de Gini e a entropia são importantes para entender a impureza de um nó em uma árvore de decisão. Certifique-se de entender como essas métricas são calculadas e quando usá-las.

- Se a sua variável de resposta é binária ou multinomial, isso irá determinar o tipo de árvore de decisão que você deve usar. No caso de uma variável de resposta multinomial, uma árvore de classificação seria apropriada.

Construa a classificação multinominal

- Ao construir uma árvore de decisão, é importante considerar a profundidade da árvore. Uma árvore muito profunda pode levar ao sobreajuste, onde a árvore se ajusta demais aos dados de treinamento e não generaliza bem para novos dados.
- O custo de complexidade é uma ferramenta útil para podar a árvore de decisão e evitar o sobreajuste. Ao construir várias árvores de decisão para diferentes valores de custo de complexidade, você pode selecionar a melhor árvore com base na acurácia na base de treino e teste.
- Após construir a árvore de decisão, é importante avaliar sua qualidade. Uma maneira de fazer isso é usando uma matriz de confusão, que mostra o número de previsões corretas e incorretas para cada classe.

Conheça o cross-validation

- Utilize a técnica de cross-validation para avaliar a eficácia de seus modelos de machine learning. Isso ajudará a entender como o resultado do modelo pode ser generalizado para uma população mais ampla.
- Divida seus dados em conjuntos de treino, validação e teste. O conjunto de treino é usado para treinar o algoritmo, o conjunto de validação é usado para testar diferentes hiperparâmetros do algoritmo, e o conjunto de teste é usado para avaliar o desempenho do modelo.
- Sempre questione e explore maneiras de obter uma métrica mais confiável para a acurácia do modelo. A escolha do conjunto de teste pode afetar essa acurácia, então é importante considerar isso ao avaliar o desempenho do modelo.

Conheça o cross-validation

- Esteja ciente de que a performance do modelo pode cair um pouco quando é avaliado em um conjunto de teste independente, após ter sido otimizado no conjunto de validação. Isso é normal e é uma das razões pelas quais a cross-validation é tão importante.

- Experimente mudar a semente do gerador de números aleatórios para ver como isso afeta a acurácia do modelo. Isso pode ajudar a entender a variabilidade do modelo.

Analise os tipos de cross-validation

- Se você tem um conjunto de dados pequeno, pode considerar o uso de métodos exaustivos, como o "leave-one-out" ou o "leave-pair-out". No entanto, tenha em mente que esses métodos podem ser computacionalmente intensivos.

- Se você deseja treinar o modelo e avaliar seu desempenho ao mesmo tempo, considere o uso de métodos hierárquicos, como o "k-fold com holdout" e o "nested k-fold".

- Para conjuntos de dados maiores, os métodos não exaustivos, como o "k-fold" e a "subamostragem sequencial", podem ser mais apropriados. Esses métodos são menos intensivos computacionalmente e ainda fornecem uma boa estimativa do desempenho do modelo.

Analise os tipos de cross-validation

- Independentemente do método de validação cruzada que você escolher, lembre-se de que o objetivo é obter uma estimativa imparcial do desempenho do modelo. Portanto, evite ajustar demais o modelo aos dados de treinamento, pois isso pode levar a um desempenho pobre em dados não vistos.

- Finalmente, lembre-se de que a validação cruzada é apenas uma parte do processo de modelagem. Você também deve considerar outras técnicas, como a regularização, para evitar o sobreajuste e melhorar o desempenho do modelo.

Conheça o K-fold

- Explore diferentes hiperparâmetros para melhorar o desempenho do seu modelo. Você pode fazer isso usando a função 'C v' e alimentando-a com um dicionário que contém as possibilidades de parâmetros a serem variados.

- Após a execução da função 'C v', explore o objeto 'grid' retornado. Este objeto contém um resumo de todos os modelos treinados, incluindo a acurácia de cada um. Use essas informações para selecionar o melhor modelo.

Conheça o K-fold

- Após selecionar o melhor modelo com base na acurácia, treine o modelo final usando a base de treino inteira e a melhor configuração de hiperparâmetros.

- Avalie o modelo final usando o conjunto de teste. Isso fornecerá uma avaliação mais realista do desempenho do modelo.

- Lembre-se de que, embora o K-Fold Cross Validation possa ser computacionalmente exigente, ele permite uma avaliação mais robusta e confiável do modelo.

escola britânica de artes criativas & tecnologia

Bons estudos!