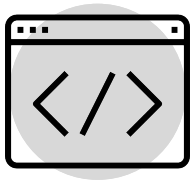
escola britânica de artes criativas & tecnologia

Profissão: Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

Árvores de regressão



- Estabeleça associações entre variáveis
- Medidas de impureza com resposta contínua
- Busque a melhor variável/algoritmo
- Visualize a árvore
- Realize a pós-poda

Estabeleça associações entre variáveis

- Utilize o conceito de erro quadrático médio (MSE) para medir a impureza. O MSE é calculado subtraindo a previsão do valor real e elevando o resultado ao quadrado. O objetivo é minimizar o MSE para que as previsões estejam o mais próximo possível dos valores reais.
- Considere também o uso do erro absoluto médio, que é semelhante ao MSE, mas usa o valor absoluto em vez do quadrado. No entanto, lembre-se de que o MSE é mais popular porque tem propriedades matemáticas interessantes e é mais rápido de calcular.
- Sempre busque por quebras que sejam significativas para o seu modelo. Uma quebra que envolve um número muito pequeno de observações pode não ser útil.

Medidas de impureza com resposta contínua

- Ao usar árvores de decisão, experimente diferentes divisões para encontrar a melhor. Não assuma que a primeira divisão que você tentar será a melhor.
- Lembre-se de que o algoritmo de árvore de decisão fará cálculos para todas as variáveis na base de dados. Não se limite a apenas uma variável ao tentar encontrar a melhor divisão.
- Ao calcular o erro quadrático médio, lembre-se de que a média pode mudar dependendo da divisão que está sendo considerada. Certifique-se de recalcular a média para cada divisão.
- A melhor divisão é aquela que resulta no menor erro quadrático médio. Não se esqueça de comparar os erros quadráticos médios de todas as divisões antes de decidir qual é a melhor.

Busque a melhor variável e algoritmo

- Ao treinar um modelo, é importante experimentar diferentes parâmetros. No caso de uma árvore de regressão, isso pode envolver a experimentação com diferentes profundidades de árvore.

- A análise de dados é um processo contínuo. Mesmo após construir e avaliar um modelo, é importante continuar a análise para entender melhor os dados e o modelo.

- Após treinar o modelo, é importante avaliar sua performance. Uma maneira de fazer isso é usando o coeficiente de determinação, ou R quadrado. No entanto, é importante lembrar que um R quadrado mais alto não significa necessariamente que o modelo é melhor. É necessário analisar mais detalhadamente para determinar a verdadeira qualidade do modelo.

Visualize a árvore

- Ao trabalhar com árvores de decisão, é importante definir parâmetros como a profundidade máxima da árvore e o número mínimo de observações em uma folha. Esses parâmetros podem ter um impacto significativo no desempenho da árvore.

- Árvores muito complicadas podem resultar em um R quadrado menor. Portanto, é importante limitar a profundidade máxima da árvore, especialmente quando se trabalha com uma grande base de dados.

- Ao variar alguns parâmetros, como a profundidade da árvore e o número mínimo de observações por folha, é possível buscar o algoritmo ótimo. Isso pode ser feito calculando o R quadrado para cada combinação e apresentando os resultados em um mapa de calor.

Realize a pós-poda

- Varie o alfa de pequeno para grande para ver como a impureza muda e construa uma nova árvore para cada C, C P. Plote a profundidade de cada árvore em função do alfa para visualizar essas mudanças.

- Calcule o erro quadrado médio (MSE) da árvore em função do alfa, usando uma lista de árvores. Isso pode ajudar a entender como o erro muda com a complexidade da árvore.

- Use o parâmetro de custo de complexidade (C, C p) para controlar a complexidade da árvore. Se um custo alto é atribuído à complexidade da árvore, ela tende a ser mais enxuta, com menos profundidade e quebras. Por outro lado, se o C C p é diminuído, a árvore é permitida a fazer todas as quebras que desejar.

escola britânica de artes criativas & tecnologia

Bons estudos!