escola britânica de artes criativas & tecnologia

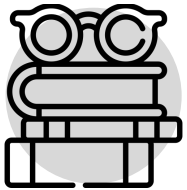
Profissão: Cientista de Dados

escola
britânica de
artes criativas
& tecnologia

GLOSSÁRIO

Árvores de regressão

- **Compreenda as árvores de regressão**
- **Encontre a melhor quebra dentro da variável**
- **Identifique a melhor variável/algoritmo**
- **Construa uma árvore de regressão**
- **Realize a pós-poda**



Dica: para encontrar rapidamente a palavra que procura aperte o comando CTRL+F e digite o termo que deseja achar.

Compreenda as árvores de regressão

Compreenda as árvores de regressão

Coeficiente de determinação (R-quadrado)

É uma medida de quão bem o modelo se ajusta aos dados. O R-quadrado varia de zero a um, onde zero significa que o modelo não explica nada do fenômeno e um significa que o modelo explica totalmente o fenômeno.

Erro Quadrático Médio (MSE)

É uma maneira de medir a impureza. O MSE é calculado subtraindo a previsão (neste caso, a média da gorjeta) do valor real e elevando o resultado ao quadrado. Isso é feito para garantir que o erro seja sempre positivo. O objetivo é minimizar o MSE para que as previsões estejam o mais próximo possível dos valores reais.

Compreenda as árvores de regressão

- **Impureza**
  É uma medida de quão bem um algoritmo pode decidir a melhor quebra em árvores de regressão. A impureza é traduzida em uma medida numérica a partir de diferenças visuais.

- **Erro Absoluto Médio (MAE)**
  É uma medida de erro que é semelhante ao MSE, mas usa o valor absoluto em vez do quadrado.

- **Quebra**
  É um termo usado em árvores de regressão para descrever o ponto onde o algoritmo decide dividir os dados. O objetivo é encontrar as quebras que resultam em maior pureza (ou menor impureza).

Encontre a melhor quebra dentro da variável

Encontre a melhor quebra dentro da variável

- **Profundidade Máxima da Árvore**
  É um parâmetro que limita o número máximo de níveis que a árvore de decisão pode ter. É usado para prevenir o overfitting.

- **Regra de Parada**
  É a condição que determina quando o algoritmo de árvore de regressão deve parar de dividir os dados. Está associada aos parâmetros definidos na árvore, como o número mínimo de observações por folha e a profundidade máxima da árvore.

Identifique a melhor variável/algoritmo

Identifique a melhor variável/algoritmo

- **Função 'Fit'**
  Método usado para treinar um modelo de aprendizado de máquina.

- **Método 'Score'**
  Método usado para avaliar a qualidade de um modelo de aprendizado de máquina.

- **Função 'test SP'**
  Função usada para dividir os dados em conjuntos de treinamento e teste.

Identifique a melhor variável/algoritmo

- **Variáveis Contínuas**
  São variáveis que podem assumir qualquer valor dentro de um intervalo específico.

- **Variável Resposta**
  É a variável que se deseja prever em um modelo de aprendizado de máquina.

- **Variáveis Explicativas**
  São as variáveis que são usadas para prever a variável resposta em um modelo de aprendizado de máquina. No contexto desta aula, são mantidas na base de dados para serem usadas na construção da árvore de regressão.

Construa uma árvore de regressão

Construa uma árvore de regressão

Pré-poda e Pós-poda

São técnicas usadas para evitar o sobreajuste em árvores de decisão. A pré-poda interrompe o crescimento da árvore antes que ela se torne perfeitamente ajustada aos dados de treinamento. A pós-poda permite que a árvore cresça completamente e, em seguida, poda a árvore para melhorar a capacidade de generalização.

Algoritmo Ótimo

É o algoritmo que fornece a melhor solução para um problema específico. No contexto desta aula, refere-se ao processo de encontrar os melhores parâmetros para a pré-poda da árvore de decisão.

escola britânica de artes criativas & tecnologia

Realize a pós-poda

Realize a pós-poda

Alfa

É um parâmetro que é variado de pequeno para grande para ver como a impureza muda em uma árvore de decisão.

Base de Testes

É um conjunto de dados separado usado para avaliar o desempenho de um modelo de aprendizado de máquina.

Realize a pós-poda

Base de Treinamento

É o conjunto de dados usado para treinar um modelo de aprendizado de máquina.

C, C p, ou Parâmetro de Custo de Complexidade

É uma ferramenta usada na pós-poda de árvores de decisão. Um custo alto atribuído à complexidade da árvore resulta em uma árvore mais enxuta, com menos profundidade e quebras.

Realize a pós-poda

Gradient Boosting

É uma técnica de aprendizado de máquina que usa árvores de decisão e é baseada no princípio de melhorar os erros de previsão de um modelo anterior.

Random Forests

É uma técnica de aprendizado de máquina que usa múltiplas árvores de decisão para fazer previsões.

Realize a pós-poda

Ruído

É a variabilidade específica da base de treinamento que não deve ser aprendida por um modelo de aprendizado de máquina.

Variabilidade

É a medida de quanto os dados em um conjunto de dados variam.

escola britânica de artes criativas & tecnologia

Bons estudos!