escola britânica de artes criativas & tecnologia

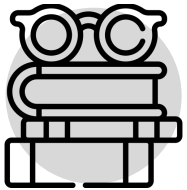
Profissão: Cientista de Dados

escola
britânica de
artes criativas
& tecnologia

GLOSSÁRIO

Combinação de modelos I

- **Conheça técnicas de ensemble - Bagging**
- **Conheça o Random Forest**
- **Compreenda Random Forest - Hyperparâmetros**
- **Implemente Random Forest em Python**
- **Ajuste os hiperparâmetros**



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

Conheça técnicas de esemble – Bagging

Conheça técnicas de esemble - Bagging

Agregação

Processo usado para determinar a resposta final do Bagging. No caso de um problema de classificação binária, a classe com a maioria dos votos dos modelos é selecionada. No caso de um problema de regressão, a média de todos os resultados dos modelos é calculada e usada como resultado final.

Amostragem com reposição

Método usado para criar um novo conjunto de dados a partir do original, com a mesma quantidade de linhas, mas com a possibilidade de repetição de linhas. É comparado com um jogo de bingo, onde a bola é devolvida ao pote após cada sorteio, permitindo a possibilidade de ser sorteada novamente.

Conheça técnicas de esemble - Bagging

Bagging

Técnica de ensemble muito popular e frequentemente questionada em processos seletivos. Envolve a criação de vários conjuntos de dados através de amostragem com reposição, a criação de um modelo para cada conjunto de dados, e a agregação dos resultados dos modelos para determinar a resposta final.

Conheça o Random Forest

Conheça o Random Forest

Overfitting

É um problema comum em aprendizado de máquina onde um modelo é treinado tão bem nos dados de treinamento que ele não consegue generalizar bem para novos dados. Isso geralmente ocorre quando o modelo é muito complexo e capta ruído ou detalhes irrelevantes nos dados de treinamento.

Random Forest

É um método de aprendizado de máquina que é uma extensão do Bagging. Ele seleciona tanto linhas quanto colunas ao criar subconjuntos de dados. Cada subconjunto é usado para construir uma árvore de decisão e a previsão final é feita por votação. O Random Forest é mais robusto e menos correlacionado do que o Bagging e é mais resistente ao overfitting.

Compreenda Random Forest - Hyperparâmetros

Compreenda Random Forest - Hyperparâmetros

GridSearchCV

Ferramenta que permite testar todas as combinações possíveis de parâmetros em um algoritmo de aprendizado de máquina para encontrar a melhor configuração.

Tunin

Processo de ajuste dos parâmetros de um algoritmo de aprendizado de máquina para melhorar seu desempenho.

Implemente Random Forest em Python

Implemente Random Forest em Python

- **Curva ROC**
  Gráfico que ilustra o desempenho diagnóstico de um sistema classificador binário à medida que o limiar de discriminação varia.

- **Grid Forest**
  Variação do algoritmo Random Forest que utiliza uma busca em grade (grid search) para otimizar os hiperparâmetros do modelo.

- **KS (Kolmogorov-Smirnov)**
  Teste estatístico usado para comparar uma amostra com uma distribuição de probabilidade ou duas amostras entre si.

Ajuste os hiperparâmetros

Ajuste os hiperparâmetros

Grid Search

É uma técnica para testar várias combinações de hiperparâmetros e encontrar a melhor. Utiliza-se a função GridSearchCV do pacote sklearn para realizar essa tarefa de forma mais eficiente.

Gini

É uma métrica utilizada para avaliar a qualidade de um modelo de classificação. Quanto maior o valor de Gini, melhor é o modelo.

Hiperparâmetros

São parâmetros que não são aprendidos durante o treinamento do modelo, mas são definidos pelo cientista de dados. Exemplos incluem o número de árvores em uma Random Forest, a profundidade máxima das árvores e o número mínimo de observações nas folhas.

escola britânica de artes criativas & tecnologia

Bons estudos!