

Relatório

Modelos de Aprendizado de Máquina para Tarefas de Regressão

Dataset1. Conjunto de Dados Sintético de Preços de Casas

1: Criar um Conjunto de Dados Sintético

Objetivo: Gerar um conjunto de dados sintético para prever preços de casas com base no número de quartos, metragem quadrada e idade da casa.

Geração de Dados:

- Criados 1.000 amostras.
- As características incluíram o número de quartos (inteiro aleatório entre 1 e 5), metragem quadrada (inteiro aleatório entre 500 e 3500) e a idade da casa (inteiro aleatório entre 0 e 100).
- Adicionado ruído para simular variações reais nos dados.
- Calculados os preços das casas usando uma combinação linear das características com o ruído adicionado.

2: Pré-processar e Dividir os Dados

Seleção de Características: Definidas as características (`bedrooms` , `square_footage` , e `age`) e a variável alvo (`price`).

Divisão de Dados: Dividido o conjunto de dados em conjuntos de treinamento (80%) e teste (20%).

3: Treinar um Modelo

Seleção de Modelo: Escolhida a Regressão Linear devido à simplicidade e interpretabilidade para uma tarefa de regressão sintética.

Treinamento: Utilizados os dados de treinamento para ajustar o modelo de Regressão Linear.

4: Avaliar o Modelo

Previsões: Realizadas previsões no conjunto de teste.

Métricas de Avaliação: Calculado o Erro Quadrático Médio (MSE) e o R-quadrado (R^2) para avaliar o desempenho do modelo.

Dataset 2. Conjunto de Dados "House Prices - Advanced Regression Techniques" (Kaggle)

1: Baixar e Carregar os Dados

Objetivo: Prever preços de casas usando dados reais do conjunto de dados "House Prices - Advanced Regression Techniques" do Kaggle.

Carregamento de Dados: Carregado o conjunto de dados a partir do arquivo CSV fornecido.

2: Pré-processar os Dados

Tratamento de Valores Ausentes: Imputados valores ausentes usando a estratégia da mediana.

Seleção de Características: Removidas colunas não numéricas e aquelas com muitos valores ausentes para simplificar a análise.

Divisão de Dados: Dividido o conjunto de dados em conjuntos de treinamento (80%) e teste (20%).

Padronização: Padronizadas as características numéricas para garantir que todas as características contribuam igualmente para o modelo.

3: Treinar um Modelo

Seleção de Modelo: Escolhida a Regressão Linear pela sua interpretabilidade e adequação para tarefas de regressão.

Treinamento: Treinado o modelo com os dados de treinamento padronizados.

4: Avaliar o Modelo

Previsões: Geradas previsões no conjunto de teste.

Métricas de Avaliação: Avaliado o desempenho do modelo usando o Erro Quadrático Médio (MSE) e o R-quadrado (R^2).

Dataset 3. Conjunto de Dados de Habitação de Boston (Kaggle)

1: Baixar e Carregar os Dados

Objetivo: Prever preços de casas usando o conjunto de dados de Habitação de Boston do Kaggle.

Carregamento de Dados: Carregado o conjunto de dados a partir do arquivo CSV fornecido.

2: Pré-processar os Dados

Tratamento de Valores Ausentes: Imputados valores ausentes usando a estratégia da mediana.

Seleção de Características: Utilizadas todas as características numéricas, exceto a variável alvo (`medv`).

Divisão de Dados: Dividido o conjunto de dados em conjuntos de treinamento (80%) e teste (20%).

Padronização: Padronizadas as características numéricas para uniformidade.

3: Treinar um Modelo

Seleção de Modelo: Seleccionada a Regressão Linear para modelar a relação entre características e preços de casas.

Treinamento: Ajustado o modelo usando os dados de treinamento.

4: Avaliar o Modelo

Previsões: Previstas os preços das casas no conjunto de teste.

Métricas de Avaliação: Avaliado o modelo usando as métricas de Erro Quadrático Médio (MSE) e R-quadrado (R^2).

Conclusão

Para cada conjunto de dados, segui uma abordagem estruturada para desenvolver e avaliar modelos de aprendizado de máquina:

1. Coleta/Geração de Dados: Obtidos ou gerados os conjuntos de dados.
2. Pré-processamento de Dados: Tratamento de valores ausentes, seleção de características relevantes e padronização dos dados.
3. Divisão de Dados: Divisão dos dados em conjuntos de treinamento e teste para avaliar o desempenho do modelo em dados não vistos.
4. Treinamento do Modelo: Treinado um modelo de Regressão Linear usando os dados de treinamento.
5. Avaliação do Modelo: Avaliado o modelo usando as métricas de Erro Quadrático Médio (MSE) e R-quadrado (R^2) para medir o desempenho.

- ❖ Essa abordagem metódica garante que os modelos sejam robustos, interpretáveis e generalizem bem para novos dados. Seguindo esses passos, podemos construir modelos preditivos eficazes para diversas tarefas de regressão.