Predicting Boston Housing Prices

REVISÃO | HISTORY

Meets Specifications

Exploração dos dados

✓ **Todas as estatísticas pedidas foram calculadas corretamente para o conjunto de dados de imóveis de Boston.**
O aluno utilizou corretamente as funções da biblioteca NumPy para obter esses resultados.

Boa! Usou o numpy corretamente. A maneira de fazer com o pandas seria: `prices.mean()`, `prices.max()` e etc.

✓ **O aluno justifica corretamente como cada atributo se correlaciona com um aumento ou diminuição na variável alvo.**

Desenvolvendo um modelo

✓ **O aluno identifica corretamente se o modelo hipotético captura a variação da variável-alvo, baseado no coeficiente de determinação (R2). O código da métrica de desempenho está corretamente implementado.**

✓ **O aluno dá uma razão válida para separar o conjunto de dados entre subconjuntos de treinamento e teste. O código de divisão em subconjuntos é corretamente implementado.**

Analisando o desempenho do modelo

✓ **O aluno identifica corretamente a tendência das curvas de treinamento e teste de acordo com o aumento do número de pontos. Também é discutido se o aumento do número de pontos beneficia o modelo.**

✓ **O aluno identifica corretamente se o modelo tem problemas de viés ou variância quando sua profundidade máxima é igual a 1 e 10, justificando através do gráfico de curvas de complexidade.**

✓ **O aluno escolheu um modelo ótimo de melhor suposição e justifica razoavelmente, usando o gráfico de complexidade do modelo.**

Avaliando o desempenho do modelo

✓ **O aluno descreve corretamente o método de busca em matriz e como ele pode ser aplicado a um algoritmo de aprendizagem.**

O Grid-Search consiste em uma busca simples pela melhor combinação de hiper-parâmetros para um determinado modelo. Hiper parâmetros são aqueles que precisamos definir antes de aprender com os dados, exemplos seriam:

- Random Forest: profundidade máxima, número de estimadores e máximo número de features a ser considerada em um split, entre outros.
- Regressão logística com regularização: o valor da penalização (L1, L2 ou ambas) e se será utilizado um intercepto ou não.

O grid-search usará a combinação de um conjunto de valores passado para cada um dos parâmetros e treinará um modelo com esse conjunto. Então, ele irá avaliar o quão bom é o modelo com aquela combinação. Tendo testado uma série de diferentes combinações de hiper-parâmetros e tendo avaliado todas, escolheremos como melhor a que possui o maior valor para a métrica escolhida (pode ser F1,

Accuracy, Precision e etc para classificação, ou RMSE, MSE e etc para um problema de regressão).

Chama-se Grid por causa dessa combinação de valores de diferentes hiper parâmetros para formar cada conjunto que será testado.

✓ **O aluno descreve corretamente o método *k-fold* de validação cruzada e discute os benefícios de sua aplicação quando usado com a busca em matriz para otimizar um modelo.**

O K-fold é uma técnica de validação, ou seja, de avaliação do modelo (embora possa ser usado para gerar previsões também). O conjunto de treino é dividido em k subconjuntos e o treinamento do modelo é realizado com k-1 desses conjuntos, enquanto a avaliação é feita no conjunto restante. Para que seja uma medida robusta, esse processo será repetido k vezes, usando um dos k subconjuntos diferentes em cada uma das k vezes. Finalmente, nós podemos tirar a média da métrica de avaliação usada para determinar quão bom ficou o modelo. Também podemos tirar o desvio padrão para saber quão sensível está o modelo à diferentes conjuntos de treinamento.

Os benefícios de usar k-fold cross validation com grid search são:

- A validação por si só, ou seja, avaliar o modelo utilizando um conjunto não visto em treinamento. Se otimizarmos os hiper-parâmetros usando uma métrica de avaliação no conjunto de teste, haveria um "leak" de informação do teste para o treino, o que não é bom, pois acaba superestimando a performance do modelo em dados não vistos. Portanto, validar usando k-fold já nos faz evitar esse cenário;
- O k-fold produz uma métrica mais robusta pelo fato de tirar a média de diferentes conjuntos de validação. Como são usados todos os dados de treino como validação em algum momento, temos uma idéia melhor da robustez do modelo;
- É uma boa alternativa para quando não se tem uma grande quantidade de dados e não poderia-se fazer um simples split entre treino, validação e teste com conjuntos suficientemente grandes.

✓ **O código da função `fit_model` foi corretamente implementado.**

✓ **O aluno identifica corretamente o modelo ótimo e o compara à sua resposta anterior.**

✓ **O aluno relata o preço de venda para os três clientes listados na tabela. A discussão sobre os preços serem razoáveis leva em consideração os dados e as estatísticas descritivas calculadas anteriormente.**

✓ **O aluno discute a fundo se o modelo deve ou não ser usado no mundo real.**

BAIXAR PROJETO

RETORNAR