Predicting Boston Housing Prices

REVISÃO | HISTORY

Requires Changes

6 ESPECIFICAÇÕES NECESSITAM DE MUDANÇAS

Caro estudante,
muito bom para a primeira submissão :D
Existem alguns pontos que gostaria que você olhasse e alterasse, mas nada muito grande.
Abraço,
Ricardo

Exploração dos dados

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

Sua intuição parece correta, mas é só baseada em palpites, certo?
Que tal plotar um gráfico para entender a correlação (exemplo, price x RM) para confirmar isso. Eu sugiro que você plote uma feature por vez ou crie 3 gráficos, uma vez que o máximo/mínimo deles são bem diferentes. Você pode brincar com o código abaixo (que é bem simples):

```
import matplotlib.pyplot as plt
plt.plot(prices, features['RM'], 'o', alpha=0.7)
# plt.plot(prices, features['LSTAT'], 'x', alpha=0.7)
# plt.plot(prices, features['PTRATIO'], '^', alpha=0.7)
plt.show()
```

↻ **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.

Certo, está funcionando, mais nós realmente queremos que você use o Numpy para calcular os valores. Tenho certeza que você pode arrumar isso sem dificuldades. Você pode checar a documentação do Numpy se precisar, exemplo: https://docs.scipy.org/doc/numpy-1.10.4/reference/generated/numpy.amax.html

Desenvolvendo um modelo

✓ **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.**

Bom trabalho! Você atribuiu um `random_state`, o que nos permite reproduzir o experimento. Além disso sua explicação é bem razoável.
Usando o conjunto de teste nos permite avaliar a performance do modelo em dados não vistos (então podemos tentar verificar se ele está realmente generalizando o problema).

✓ **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.**

Certo, bom trabalho aqui! O R-squared é uma medida estatística do quão bem o modelo encaixa nos dados. É também conhecido como coeficiente de determinação. Podemos dizer que o R-squared é a porcentagem da variacão que é explicada pelo modelo (ou seja, é sempre entre 0 e 100%).

- 0% indica que o modelo não explica a variabilidade dos dados ao redor da média
- 100% indica que o modelo explica toda a variabilidade

Então, maior melhor. Nesse caso, pode ser explicado 92% da variabilidade então os modelo deve prever bem os dados

bem os dados.

Analisando o desempenho do modelo

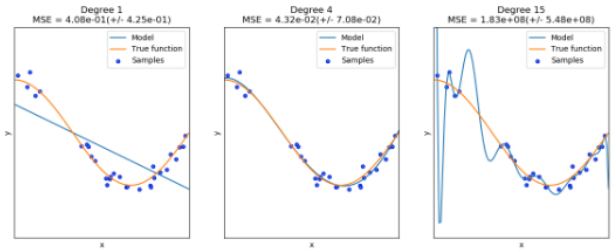
O aluno identifica corretamente a tendência das curvas de treinamento e teste de acordo com o aumento do número de pontos. Também é discutido se o aumento do número de pontos beneficia o modelo.

Sua observação está correta. Adicionar mais amostras para treinamento porvavelmente não melhora o modelo.

Além disso, o `max_depth=10` é um caso curioso pois geralmente olhamos o score de treinamento e pensamos "Nossa! O score está ótimo! Está realmente aprendendo algo!", mas se não olharmos o score do teste podemos falhar miseravelmente :(

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.

Perfeito. Esses conceitos são realmente importantes e seria impossível identificar o over/underfitting sem os conjuntos de treinamento e teste. Eu gosto muito da documentação do Sklearn (e costumo visitar ela com uma certa frequência). really like the Sklearn documentation (and I check it almost every day).. ela tem uns exemplos ótimos! Gostaria que você desse uma olhada nesse (que é relacionado ao assunto):

> Underfitting vs. Overfitting
>
> 
>

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

Com `max_depth=1` e `max_depth=3` o modelo converge, mas o `max_model=3` tem um score mais alto. Com as profundidades mas altas, eles nem convergem (e realmente ocorre o overfitting). Talvez, com MUITAS amostras a mais, ele teria resultados melhores com o `max_depth=6` e `max_depth=10`, mas só testando para ter certeza.

Avaliando o desempenho do modelo

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.

Você realmente entendeu a ideia do k-fold, mas qual é a vantagem dela para otimizar o modelo (com relação gridsearch, por exemplo)? Como é escolhido o melhor modelo? Recomendo que você de uma olhada nessa página para te ajudar :)
https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

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

Certo, você está no caminho certo.. mas o que é o grid? Como funciona os testes dos parâmetros? Como escolhemos o melhor? Gostaria que você elaborasse um pouco mais sua resposta :)

O código da função `fit_model` foi corretamente implementado.

Opa, muito bom ter trabalahdo com o `random_state` :D
Uma dica aqui para deixar seu código mays pythonico: você poderia uasr `range(1, 10)`. Aqui não tem problema, mas e se você tivesse muitos hyper-parâmetros e ranges maiores?

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

Certo, mas você esqueceu de comparar com seu palpite da questão 6 :)

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.

Na verdade você deveria sugerir o output do modelo, principalmente por estar colocando uma faixa muito grande. A ideia daqui é dizer se o preço que o modelo preveu para essas casas faz sentido (pela quantidade de quartos, professores e classe baixa). O ideal seria comparar elas.

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

Concordo com quase tudo que você disse, exceto pelo começo. Poder, o modelo pode ser usado.. mas você nao deveria usa-lo cegamente.. A variação que foi encontrada é muito grande, então o modelo não é robusto o suficiente. Além disso, só as correções não são suficientes para garantir o valor correto dos imóveis (MUITA coisa mudou)... pense que o comércio, transporte público e etc. mudaram. Gostaria que você trabalhasse um pouco mais nessa resposta :)

REENVIAR PROJETO

BAIXAR PROJETO

Melhores práticas para sua resubmissão do projeto

Ben compartilha 5 dicas úteis para a revisão ressubmissão do seu projeto.

Assistir Vídeo (3:01)

RETORNAR