Sessão 04

WiDS Recife Live Coding, 11/01/2020

Como serão os live codings?

- Sessões ao-vivo todos os sábados das 14h às 15h
- Código e slides serão disponibilizados no nosso site
- O objetivo é treinar para participar do Datathon em 2020

Recapitulando!

Recapitulando

- Usamos o Pandas para explorar os dados
- Usamos o scikit-learn para selecionar as melhores features e treinar o modelo com o algoritmo de Árvore de Decisão
- Conhecemos algumas métricas para avaliar a performance de um modelo
- Avaliamos o modelo para identificar “onde ele está errando”
- Slides da sessão passada: https://github.com/widsrecife/live-coding

Começando!

Roteiro

- Entender o que é viés e o que é variância e como isso impacta na avaliação dos modelos
- Conhecer formas de particionar os dados para avaliar os modelos
- Usar curvas de aprendizagem para avaliar o viés e a variância dos modelos

O problema

- Nosso conjunto de dados é composto por atributos de vestidos
- Queremos treinar um modelo onde a gente envie os atributos do vestido e ele diga **qual a melhor época do ano para usá-lo**

Etapas para resolver o problema

1. Importar os dados
2. Explorar os dados
3. Treinar o modelo com o conjunto de treinamento
4. Avaliar o modelo com o conjunto de testes

5. Avaliar o modelo

- 5. Avaliar o modelo
  - Duas fontes de erros comuns: viés e variância (bias e variance em Inglês)

Principais fontes de erro nos modelos de predição

- Viés (bias)
- Variância (variance)
- Erro irredutível (Irreducible Error)

Viés (bias)

- O que é
  - É a incapacidade do modelo em classificar corretamente os dados, quando o modelo não consegue identificar a relação entre as features de entrada e as classes de saída
- Quando acontece
  - Quando o modelo é simples demais para aprender como modelar os dados (**underfitting**)
- Exemplo
  - Criar uma árvore com max_depth == 2

Variância (variance)

- O que é
  - É quando o modelo classifica bem um conjunto de dados mas se comporta de forma ruim com novos dados
- Quando acontece
  - Quando o modelo “aprende muito” os dados do conjunto de treinamento, fazendo com que ele não generalize bem (**overfitting**)
- Exemplo
  - Criar uma árvore de decisão para cada instância dos dados da mesma estação

“Erro irredutível” (irreducible error)

- O que é
    - Um erro atribuído à aleatoriedade inerente aos dados

Complexidade do modelo

- Tanto o viés quanto a variância estão relacionados à complexidade do modelo
  - Modelo muito simples -> viés alto (underfitting)
  - Modelo muito complexo -> variância alta (overfitting)
- É **impossível** reduzir o viés sem aumentar a variância e é **impossível** reduzir a variância sem aumentar o viés

Como os dados influenciam na avaliação

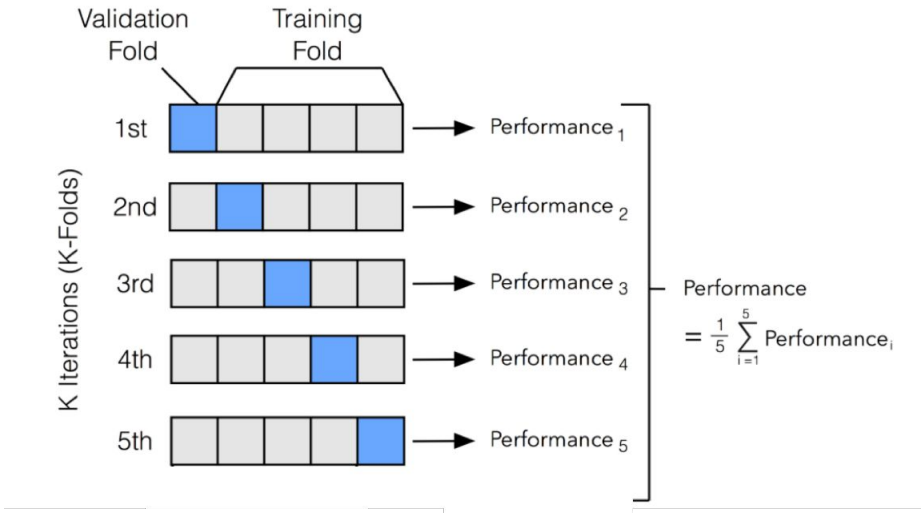
Avaliando a capacidade do modelo generalizar

- Normalmente temos poucos dados disponíveis, e se usarmos o conjunto de testes para avaliar e melhorar o modelo corremos o risco de aumentar a variância
- Uma técnica útil para avaliar a capacidade de generalização do modelo é a **validação cruzada**
    - O objetivo é tentar estimar quão bom um modelo é na prática (na prática: quando a gente sair do cenário de testes e for para o cenário real)

Validação cruzada

- Consiste em particionar os dados em conjuntos mutuamente exclusivos
- As três formas mais utilizadas de se fazer isso são:
  - Método holdout (separar os dados em treinamento e teste)
  - Método k-fold (dividir os dados em k-conjuntos mutuamente exclusivos)
  - Método leave-one-out (erro calculado para cada dado)

K-fold



Exemplo onde k = 5

Fonte: http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

Stratified k-fold

- Preserva as porcentagens das instâncias dos dados para cada classe

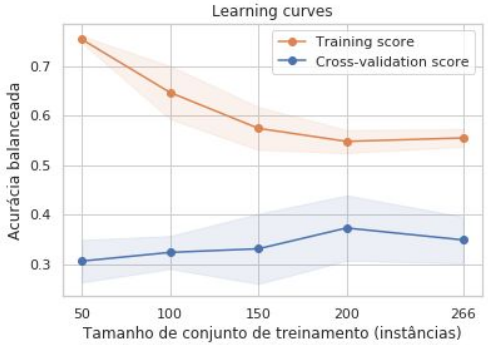
Formas de avaliar o modelo quanto ao viés e a variância

Identificando viés e variância nos modelos

- Existem fórmulas teóricas é difícil calcular os valores reais, então uma forma mais simples é estimar os erros de forma empírica com **curvas de aprendizagem** de **curvas de validação** [1]
- Vamos plotar a **curva de aprendizagem**

Curva de aprendizagem

- Mostra a performance do modelo ao variar o número de instâncias do conjunto de treinamento
  - Curva de treinamento: score do conjunto de treinamento
  - Curva de validação: score no conjunto da validação cruzada



Curva de aprendizagem - Diagnosticando

- As duas curvas convergindo para um valor baixo: provavelmente o modelo não se beneficia ao adicionar mais dados
- As duas curvas estão perto: underfitting
- As duas curvas estão longe: overfitting

Obrigada!

E até semana que vem!

—

Referências

- [1] https://www.packtpub.com/data/hands-on-ensemble-learning-with-python
- [2] https://pt.wikipedia.org/wiki/Valida%C3%A7%C3%A3o_cruzada
- [3] https://scikit-learn.org/stable/modules/learning_curve.html
- [4] https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/
- [5] https://github.com/jakevdp/sklearn_tutorial/blob/master/notebooks/05-Validation.ipynb