📙

A estratégia do Treino-Teste

Ciclo	Ciclo 05: As garantias de aprendizado
# Aula	27
Created	@February 16, 2023 2:25 PM
Done	☐
Ready	☑

Objetivo da Aula:

- ☐ A estratégia do Treino-Teste
- ☐ O problema da separação Treino-Teste
- ☐ A solução de Joaquim
- ☐ Resumo
- ☐ Próxima Aula

Conteúdo:

▼ 1. A estratégia do Treino-Teste

Existe uma estratégia melhor para medir a capacidade de generalização de um algoritmo de Machine Learning do que colocá-lo diretamente em Produção.

Essa estratégia consiste em separar um conjunto de dados em 2 subconjuntos: Treinamento e Teste.

▼ 1.1 O modo de separação

A separação dos dados em conjunto de treinamento e teste deve ser feita de maneira aleatória, mantendo a proporção original dos exemplos entre as classes. Por exemplo:

▼ 1.1.1 Conjunto de dados:

Cor	Combustível	Preço	Número de Portas	Tipo de Veículo
Azul	Gasolina	R$ 30.000	4	Carro
Vermelho	Gasolina	R$ 22.500	2	Carro
Verde	Álcool	R$ 40.000	4	Carro
Preto	Diesel	R$ 35.000	4	Carro
Prata	Gasolina	R$ 10.000	0	Moto
Branco	Gasolina	R$ 12.000	0	Moto
Amarelo	Álcool	R$ 12.500	0	Moto
Azul	Gasolina	R$ 11.000	0	Moto
Cinza	Diesel	R$ 300.000	2	Ônibus
Verde	Diesel	R$ 350.000	2	Ônibus
Azul	Diesel	R$ 290.000	2	Ônibus
Preto	Diesel	R$ 320.000	2	Ônibus
Amarelo	Diesel	R$ 280.000	2	Ônibus
Vermelho	Diesel	R$ 270.000	2	Ônibus

▼ 1.1.2 Exemplo da separação entre Treinamento e Teste

1. Conjunto de dados original (100% dos dados)
    a. 25 colunas e 10.000 linhas
    b. 60% classe A e 40% classe B
2. Conjunto de dados de treino (80% dos dados originais)
    a. 25 colunas e 8.000 linhas
    b. 60% classe A e 40% classe B
3. **Conjunto de dados de teste** (20% dos dados originais)
    a. **25 colunas e 2.000 linhas**
    b. **60% classe A e 40% classe B**

▼ 1.2 A proporção da separação

As proporções mais comuns da separação são:

1. 80% treinamento e 20% teste
2. 70% treinamento e 30% teste

▼ 2. O problema da separação Treino-Teste

Joaquim receber um segundo problema de negócio. Ele entendeu a real necessidade do time de negócio, fez um planejamento de execução das tarefas, coletou e limpou os dados, preparou os dados e treinou alguns algoritmos de Machine Learning. No final do treinamento, ele ficou na dúvida ao escolher entre dois tipos de algoritmos: Um modelo linear e um modelo polinomial.

Joaquim lembrou que não aprendeu como escolher entre dois ou mais algoritmos de Machine Learning, durante sua pós-graduação no FAPONE. E agora começou a pensar: “Como eu posso decidir entre eles?”

Depois de um certo tempo pensando, Joaquim julgou que a opção mais óbvia seria treinar ambos e comparar como eles generalizam usando o conjunto de teste. Depois de alguns testes, ele descobriu que o modelo de regressão linear generalizou melhor.

Mas antes de colocar em Produção, Joaquim gostaria de aplicar uma técnica chamada regularização para evitar um grande problema chamado overfitting. Essa técnica requer que você escolha um único valor para a regularização, dentre vários possíveis.

Joaquim, então, encontram-se no segundo problema: “Como eu escolho o valor do parâmetro de regularização? Uma opção é treinar 100 modelos diferentes usando 100 valores diferentes para esse parâmetro”.

Após uma bateria de testes, Joaquim encontrou o melhor valor do parâmetro que permite o algoritmo produzir o menor erro de generalização, apenas 5% de erro.

Joaquim dá pulos de felicidade e como a se questionar quando foi que ele ficou tão esperto, inteligência e astuto. Aquele nem parecia ele. Com toda a confiança do mundo, Joaquim publica o seu algoritmo de Regressão Linear regularizada com o melhor parâmetro em Produção, mas infelizmente ele não funciona como o esperado e produz erros de 15%.

▼ O que acabou de acontecer?

O problema é que Joaquim mediu o erro de generalização várias vezes no conjunto de teste e adaptou o modelo e os parâmetros para produzir o melhor modelo *para esse conjunto de teste em particular*. Isso significa que o modelo provavelmente não funcionará tão bem em novos dados.

▼ 3. A solução de Joaquim

A estratégia de Validação Holdout

▼ 4. Resumo

1. A criação de dois conjuntos de dados deve ser feita de maneira aleatória, mantendo as proporções originais das classes.
2. A separação em Treino-Teste não é suficiente para descobrir os melhores parâmetros de uma algoritmo de ML treinado.

▼ 5. Próxima aula

A estratégia de Validação Holdou