

12

Calculando as variáveis de treino, teste e validação

Atualmente, estamos realizando apenas as duas fases no nosso algoritmo, que são o treino e teste, porém, vimos que também a terceira fase que é escolher o algoritmo que obteve o melhor resultado e então testá-lo com dados do mundo real. Para isso precisamos primeiro desses dados do mundo real, isto é, os dados de validação. Nesse instante, faremos a seguinte distribuição dos dados para cada uma das variáveis:

- 80% para teste.
- 10% para treino.
- 10% para validação.

Primeiro altere o valor da variável `porcentagem_de_treino` para `0.8`, agora crie a `porcentagem_de_teste` com valor de `0.1`. Com os percentuais definidos, precisamos definir o tamanho de cada uma das variáveis, o tamanho das variáveis `tamanho_de_treino` e `tamanho_de_teste` é justamente o `len` de seus percentuais:

```
tamanho_de_treino = porcentagem_de_treino * len(Y)
tamanho_de_teste = porcentagem_de_teste * len(Y)
```

Porém, o tamanho de validação precisa ser a quantidade restante que não faça parte do tamanho de treino e teste, para isso, pegue todos os elementos (`len(Y)`), subtraia pelas variáveis `tamanho_de_treino` e `tamanho_de_teste` e atribua para a variável `tamanho_de_validacao`.

Com todos os tamanhos em mãos, podemos calcular as variáveis de dados e marcações de treino, teste e validação. Podemos começar pela variável de treino, lembre-se que ela precisa ser os 80% dos dados, então não há nenhuma novidade, ou seja, podemos utilizar o mesmo código que já temos:

```
treino_dados = X[:tamanho_de_treino]
treino_marcacoes = Y[:tamanho_de_treino]
```

Porém, para o teste, precisamos pegar todos os dados a partir do tamanho de treino até o tamanho de treino mais o tamanho de teste que representa justamente os próximos 10% após os 80% de treino. Porém, perceba que fazer `tamanho_de_treino + tamanho_de_teste` é uma instrução um pouco grande e que não é tão clara o motivo, ou seja, crie uma variável chamada `fim_de_treino` e atribua a ela o valor dessa operação, então a utilize no lugar dessa operação.

Por fim, precisamos pegar o restante, ou seja, os 10% que sobraram para os dados e marcações de validação. Para isso, basta apenas falar que você quer os dados a partir da variável `fim_de_treino`, pois ela representa os 90% que já foram utilizados. Cole o código abaixo.

Responda

INserir código		Formatação

