02

Datasets de treino e teste

Transcrição

Voltando com nosso projeto, iremos de fato estimar nosso modelo de regressão linear. Regressão linear nada mais é que o estudo da dependência de uma variável - em nosso caso a variável de consumo - em relação a um conjunto de variáveis explicativas. O objetivo é prever ou estimar o valor médio da variável dependente, tendo antes conhecido o valor das variáveis dependentes.

Para tanto, utilizaremos o **Scikit-learn**, uma biblioteca especializada em *machine learing* que possui um ferramental completo para esse fim. Inicialmente, separaremos nossos dados em *train* e *test* para modelar nossa regressão e testá-la. Importaremos a função `train_test_split` do sckit-learn para executar esta tarefa.

Feita a importação, precisaremos preparar nosso dataset geral, isto é, nosso dataframe `dados` . Precisaremos dividir o dataframe em series que contém a variável dependente e outro dataframe que contém as variáveis explicativas. Então criaremos:

```
y = dados['consumo']
```

COPIAR CÓDIGO

e então:

```
X = dados[['temp_max', 'chuva', 'fds']]
```

COPIAR CÓDIGO

Neste ponto, usaremos a função `train_test_split` para separar os conteúdos entre treino e teste. Essa função possui como retorno uma lista de quatro itens, e precisamos atribuir a cada elemento da lista o conteúdo das variáveis. O primeiro item da lista é o x de treino, o segundo é o x de teste, depois y de treino e y de teste. Configuraremos, ainda o `test_size` como `0.3` ,que se refere à quantidade de dados que usaremos para teste, então 30% dos dados serão selecionados de forma aleatória para a realização de testes do modelo, já o restante será para treino.

O próximo parâmetro é o `random_state` . Ao fixarmos um valor, por exempLo `2811` , teremos um conjunto de treino e de teste idênticos.

Ao final, teremos a configuração: `x_train.shape` de 225,3 ; `x_test.shape` de 110,3.