03

Conhecendo o dataset

Transcrição

Daremos inicio ao nosso projeto utilizando a regressão linear. O objetivo desta primeira aula é nos familiarizarmos com o dataset que utilizaremos ao longo do curso. É importante que conheçamos a estrutura de pastas que criamos, e você deve fazer o download de todas elas que estão disponíveis no tópico "Projeto Inicial do Treinamento".

A documento do projeto foi armazenado em "C:\Usuários\Alura\data-science". Em "data-science" encontraremos "reg-linear", que armazena as patas criadas do projeto. Na pasta "Dados" teremos o dataset `Consumo_cerveja.csv`. Ainda em "data-science" teremos a pasta "projeto", em que encontraremos o notebook `Regressão Linear.ipynb`.

Na pasta "data-science", teremos o arquivo "StartJupyter". Clicaremos sobre ele para iniciar a ferramenta. O Jupyter já será executado na pasta que estamos trabalhando. Por fim, clicaremos sobre "Projeto" e abriremos o notebook `Regressão Linear`.

Ele estará inteiramente documentado, apenas as células vazias serão preenchidas, e já contém todas as aulas que desenvolveremos ao longo do treinamento. Começaremos importando as bibliotecas básicas.

Caso você já tenha feito o curso de Pandas, encontrará algumas similaridades. Importaremos `matplotlib.pyplot` com o "apelido" `as plt`. Depois, inseriremos a function `%matplotlib inline`. Alguns sistemas precisam dessa

configuração para que o Jupyter consiga imprimir os gráficos. Em seguida, importaremos o Pandas e o Numpy.

No notebook, temos um espaço para bibliotecas opcionais. Acessaremos o portal Kaggle (http://www.kaggle.com/dongeorge/beer-consumption-sao-paulo), onde encontraremos uma grande quantidade de datasets voltados para data science.

Para este projeto faremos algo bem simples, com o tema consumo de cerveja. Teremos uma breve descrição do arquivo, e também uma distribuição de frequência de cada variável.

O objetivo do nosso projeto é estimar um modelo de machine usando a técnica de regressão linear, e averiguar os impactos das variações disponibilizadas no dataset, sobre o consumo de cerveja. Tentaremos estimar o consumo utilizando a regressão utilizando as variáveis apresentadas. Vamos conhecer cada uma delas:

`data` = dia de coleta. `temp_media` = média da temperatura ambiente registrada `temp_min` = temperatura mínima ambiente registrada `temp_max` = temperatura máxima ambiente registrada `chuva` = Precipitação(mm) `fds` = Final de Semana (1= Sim, 0=Não) `consumo` = Consumo de Cerveja (litros)

Temos uma questão: os dados do Kaggle que iremos utilizar faz uso da `,` ao invés do `.` : por exemplo `27,3` . De volta ao notebook `Regressão Linear` , na célula de "Leitura dos dados", importaremos o arquivo o seguinte arquivo - lembrando de especificar o separador:

```
dados = pd.read_csv('../Dados/Consumo_cerveja.csv', sep='
```

COPIAR CÓDIGO

Assim feito, basta escrever `dados` na próxima célula para gerarmos a visualização da tabela, composta pelas variáveis que já conhecemos e seus respectivos valores. Verificaremos o tamanho do nosso dataset ao escrever:

```
dados.shape
```

COPIAR CÓDIGO

Teremos como resultado `(365,7)`, isto é, uma dupla numérica, em que o primeiro valor corresponde ao número de linhas de registro no dataset e o segundo ao número de variáveis.