Relatório 04 - Principais Bibliotecas e Ferramentas Python para Aprendizado de Máquina

Guilherme Loan Schneider

Descrição da atividade

Esse documento contém os principais tópicos abordados de forma resumida. O detalhamento de cada função apresentada nas aulas está nos arquivos enviados no Github.

O primeiro módulo de aulas contém uma breve explicação e iniciação do ambiente a ser utilizado durante o curso, o Jupyter Notebooks. Além disso, mostra também a criação de ambientes de trabalho virtuais, onde é possível cria-los com versões diferentes de bibliotecas, interpretadores, dentre outros, como possuir o Pyhton 3.11 em sua máquina e precisar do 3.6 por possuir maior estabilidade, por exemplo.

Em seguida temos um módulo explicando a biblioteca Numpy, voltada para equações e manipulações em dados. Esse módulo possui como conteúdo os seguintes itens:

- Noções básicas de importação e utilização;
- Introdução ao Numpy, comparando as alocações tradicionais com a utilização da biblioteca em diferentes cenários;
- Arrays, vetores e matrizes, demonstrando a criação de cada um, utilizando várias formas diferentes, como np.arrange(), np.linspace(), np.zeros() (esse último voltado para matrizes), dentre outros;
- Operações com Numpy Arrays, aplicando soma, divisão, multiplicação, potenciação, resto, dentre inúmeros outros cálculos. Importante salientar que o numpy gera warnings ao invés de erros em casos de erros, como a divisão por 0, não interrompendo a execução do código.
- Por fim, há uma seção de exercícios que consistiram em percorrer os arrays, cria-los a partir da função np.arrange() utilizando passo, reorganizar um array em uma matriz (vetor de tamanho 9 passa para uma matriz de tamanho 3x3), além da aplicação de funções matemáticas como média e somatório.

Por fim, o módulo 6 explica e utiliza a biblioteca Pandas em visualização, limpeza e análise de dados. Essa biblioteca pode trabalhar com vários tipos de dados diferentes, que vão desde arquivos do Excel, até códigos HTML. Abaixo estão listados os principais conteúdos abordados nessa seção:

- Noções básicas de importação e introdução ao Pandas, abordando qual a sua finalidade e como será utilizado;
- O primeiro conteúdo abordado são as Series (similares a dicionários), no entanto, são voltados para a manipulação e visualização de dados pelo Pandas. Além disso, demonstra algumas manipulações possíveis nesse tipo de dado.
- A contextualização do que são DataFrames (conjuntos de Series), bem como a criação e fatiamento. A alocação básica de um DataFrame consiste em passar para a função, um vetor ou matriz (os dados a serem analisados), um vetor de *index* (serão as linhas do nosso DF) e por fim um vetor de columns (serão as colunas do nosso DF).
- Seleção condicional em DFs e adicionar uma nova coluna, alterando as *tags* de linhas com a função set_index. A seleção condicional consistiu em retornar valores binários, Verdadeiro ou Falso.
- Índices multiníveis, ou seja, utilizar uma hierarquia nas linhas de uma matriz, como no exemplo abaixo. É implementado a partir de dois métodos, o primeiro sendo o list(zip(INDICE_SUPERIOR, INDICE_INFERIOR)), e o segundo o MultiIndex.from_tuples(), que recebe a lista gerada pelo método anterior.

		A	B
G1	1	1.053444	1.231848
	2	1.477577	1.054774
	3	-0.082864	0.631808
G2	1	0.710484	1.564710
	2	-0.060671	1.262078
	3	-0.030929	-0.344591

- A biblioteca Pandas é utilizada também para tratamento de dados, onde existem várias formas de se preencher um valor nulo em um conjunto de dados. A primeira delas e mais simples é o dropna(), que basicamente remove todas as linhas que possuem valores nulos. A segunda seria preencher utilizando a média dos valores daquela coluna, e por fim, o preenchimento do dado utilizando a célula anterior ou posterior. As imagens abaixo ilustram o preenchimento a partir da célula anterior.

	A	B	C
0	1.0	5.0	1
1	2.0	5.0	2
2	2.0	5.0	3

	A	B	C
0	1.0	5.0	1
1	2.0	NaN	2
2	NaN	NaN	3

- As funções de GroupBy, Concatenar, Juntar e Mesclar funcionam da mesma forma que em bancos de dados. O GroupBy utilizará valores repetidos em uma tabela, o Concat junta as tabelas de tamanhos iguais (pode ser concatenado utilizando os índices de linha), o Join permite juntar tabelas de acordo com índices em comum, e utiliza também o inner/outer join. Por fim, a Mesclagem, que junta tabelas de acordo com uma coluna em comum, utilizando a ideia do inner/outer join, passando a coluna em comum da tabela.
- Operações – Resumidamente, teve-se inúmeras operações a serem utilizadas em um DataFrame, estas por sua vez estarão melhores explicadas no próprio arquivo pd_operacoes.py. Abaixo estão explícitas todas as funções abordadas nessa seção da aula:
    - unique(); nunique(); condições que retornam valores do DF; funções lambda; columns e index; sort_values(); isnull(); dropna(); fillna() e pivot_table();
- Por fim, os tipos possíveis de entrada e saída de dados da biblioteca Pandas, que permite a entrada de inúmeros arquivos, como dito anteriormente. Acredito que os principais a serem utilizados são arquivos csv, xlsx, xml e html, esses por sua vez podem estar na máquina local ou na internet. A saída de dados é possível a partir da função to_<formato_arquivo>, salvando o dataframe na máquina do usuário.

Conclusões

Este relatório apresentou uma introdução ao ambiente de trabalho Jupyter Notebooks e à criação de ambientes virtuais. O módulo sobre Numpy abrangeu operações matemáticas e manipulação de arrays, enquanto o módulo de Pandas abordou a visualização, limpeza e análise de dados com Series e DataFrames. Ferramentas para manipulação condicional, tratamento de valores nulos e operações como GroupBy e Mesclagem, foram exploradas, assim como as diversas formas de entrada e saída de dados.

Referencias

Foram utilizadas as aulas disponibilizadas no card da atividade.

Python para Data Science e Machine Learning