Relatório 4 - Principais Bibliotecas e Ferramentas Python para Aprendizado de Máquina (I)

Yuri Vacelh Zamulhak Zdebski

Descrição da atividade

Assistir aos módulos 3, 5, 6, 7 do curso Python para Datascience e Machine Learning, de Rodrigo Soares Tadewald e desenvolver uma pequena aplicação com os conhecimentos adquiridos.

Jupyter

Essa sessão é puramente introdutória, com o intuito de mostrar como utilizar o software Jupyter, plataforma que organiza o código em células, tendo a vantagem de ser extremamente didático (por separar o código, facilita ver o que cada parte faz), e a criação e utilização de ambientes virtuais, que tem como objetivo evitar conflitos de versões de bibliotecas para diferentes softwares rodando em uma mesma máquina.

Numpy

Essa seção trata da famosa biblioteca Numpy, que aborda principalmente questões de álgebra linear. Implementada em C, ela é extremamente performática sem trazer a complexidade da linguagem, sendo perfeita para trabalhar com grandes volumes de dados de forma prática e elegante.

A principal estrutura de dados da biblioteca são os arrays, que permitem a manipulação eficiente de grandes volumes de dados numéricos. Eles são homogêneos, ou seja, todos os elementos de um array possuem o mesmo tipo de dado, o que permite otimizações de memória e processamento. Arrays suportam operações vetorizadas, possibilitando a execução de operações elementares em todos os seus elementos de forma simultânea, resultando em um desempenho significativamente superior ao das listas Python tradicionais para cálculos.

Os principais pontos trazidos no curso são:

- Como criar arrays (np.eye(), np.zeros(), np.linspace(), entre outras formas citadas)
- Como manipular arrays (essa parte trata principalmente do operador de slices, que é basicamente o mesmo utilizado em listas array[início:fim:passo])
- Operações em arrays (operações como soma, multiplicação por escalar entre outras, são realizadas em cada elemento do array)

Um ponto que achei importante mencionar é que, quando um array é atribuído da seguinte forma array1 = array2, não é criada uma cópia de array2 em array1, mas sim uma referencia (ou seja, se modificarmos um, o outro será afetado), portanto caso seja necessária uma cópia, o método .copy() é utilizado.

Pandas

A última seção da atividade (módulos 6 e 7) aborda a biblioteca Pandas, projetada para manipular grandes volumes de dados organizados em tabelas de diversos formatos, como .csv e .xlsx.

A unidade básica da Pandas são as Series. As Series funcionam como dicionários, com estrutura de chave-valor, e são similares aos arrays Numpy, pois são construídas sobre a Numpy. O acesso aos elementos é feito por meio de chaves ([<key>]), e muitas operações realizadas sobre essas estruturas assemelham-se às da Numpy. No entanto, o foco principal da biblioteca reside na próxima estrutura que vamos explorar, os DataFrames.

Figura 1: DataFrame

	W	X	Y	Z
A	-0.993263	0.196800	-1.136645	0.000366
B	1.025984	-0.156598	-0.031579	0.649826
C	2.154846	-0.610259	-0.755325	-0.346419
D	0.147027	-0.479448	0.558769	1.024810
E	-0.925874	1.862864	-1.133817	0.610478

Fonte: Retirado do Curso

A figura acima ilustra o conceito de DataFrames, onde cada coluna é uma Series. A principal funcionalidade da Pandas é o manuseio de dados tabulares, tornando a estrutura de tabela essencial. Os principais tópicos abordados no curso incluem:

- Operações básicas em DataFrames
- Entrada e saída de dados
- Consultas em DataFrames
- Agrupamento de dados
- Importação de tabelas de diferentes formatos

Conclusões

As bibliotecas introduzidas neste módulo são cruciais para a geração de conhecimento em qualquer área que envolva dados, incluindo, claro, o Machine Learning. Com os conhecimentos adquiridos, realizei uma análise simples (talvez mais uma demonstração da utilização das bibliotecas do que uma análise propriamente dita) em um conjunto de dados climáticos obtido do site Kaggle.