Machine Learning Bootcamp

Lecture 4: Pandas

What is Pandas?

Pandas

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.

- It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.
- It has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language.
- It is already well on its way toward this goal.

Why pandas?

pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data need not be labeled at all to be placed into a pandas data structure

Getting Started with pandas

Installation

- pip install pandas

Import

- import pandas as pd

Getting data

- https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

Reading Data into pandas dataframe

Methods

- read_csv
- read_excel

Code snippet

```
import pandas as pd
df = pd.read_csv(“File_Path”)
df2 = pd.read_excel(“File_Path”)
```

Looking into data

Methods

- head
- tail

Code snippet

```
import pandas as pd
df = pd.read_csv(“File_Path”)
df2 = pd.read_excel(“File_Path”)
First_five_rows = df.head()
Last_five_rows = df.tail()
```

Describing data

Methods

- info
- describe
- nunique

Code snippet

```
import pandas as pd
df = pd.read_csv(“File_Path”)
Df.shape
df.columns
df.info()
df.describe()
def.nunique()
```

Accessing data

Technique

- iloc
- Through column names

Code snippet

```
import pandas as pd
df = pd.read_csv(“File_Path”)
df_Col = df[‘name_of_column’]
Df_cols = df[[‘col1’,’col2’,’col3’]]
Df_cols = df.iloc[rows,columns]
```

Dealing with missing data

Method

- isnull
- isna
- dropna
- fillna

Code snippet

```
import pandas as pd
df = pd.read_csv(“File_Path”)
df.isnull().sum().sum
df.isna().sum()
df.dropna(axis = 0, thresh = 6)
df. fillna(value=None, method=None, axis=None, inplace=False)
```

Resources

- Slides
- Video Recordings
- https://github.com/owais4321/mlbootcamp

Assignment

- https://github.com/owais4321/mlbootcamp
- Due before next class
- Submission on google forms