Machine Learning CSE4020

Prof. Ramesh Ragala

December 10, 2018

Introduction to Pandas

- Pandas is an open source, BSD-licensed library
- Pandas is a newer package built on top of NumPy
- It provides an efficient implementation of a DataFrame in Python.
- DataFrames are essentially **multidimensional arrays with attached row and column labels**, and often with heterogeneous types and/or missing data.
- It offers a convenient storage interface for labelled data.
- Pandas implements a number of powerful data operations familiar to users of both database frameworks and spreadsheet programs
- High-performance, easy-to-use data structures and data analysis tools
- Built for the Python programming language

Introduction to Pandas Series

- It is a one-dimensional array of indexed data.
- It can be created from a list or array.
- **data = pd.Series([0.25,0.5,0.75,1.0])**
- **print(data)**
- **data.values**
- **data.index** $\rightarrow$ Values and index are attributes
- **data[1]** $\rightarrow$ Check output
- **data[1:3]** $\rightarrow$ Check output

- **data = pd.Series([0.25, 0.5, 0.75, 1.0], index=['a', 'b', 'c', 'd'])**
- **print(data)**
- **data['b']**
- **data = pd.Series([0.25, 0.5, 0.75, 1.0], index=[2, 5, 3, 7])**
- **data** $\rightarrow$ Check output
- **data[5]** $\rightarrow$ Check output

- **populationdict = {'California': 38332521, 'Texas': 26448193, 'New York': 19651127, 'Florida': 19552860, 'Illinois': 12882135}**
- **population = pd.Series(populationdict)**
- **populationdict**
- Series will be created where the index is drawn from the sorted keys.
- **populationdict['Florida']**
- Series supports array-style operations such as slicing, etc.
- **populationdict['California':'Florida']**

Introduction to Pandas Series

- areadict = {'California': 423967, 'Texas': 695662, 'New York': 141297, 'Florida': 170312, 'Illinois': 149995}
- area = pd.Series(areadict)
- area → check output
- states = pd.DataFrame({'population': population, 'area': area})
- states
- states.index
- states.columns
- states.columns