Pandas Data Structures

Mike West
MACHINE LEARNING ENGINEER

Module Overview

- Series object
- The array
- Domain knowledge
- Methods and functions
- High-level data cleansing steps
- The final Wrangle

The Series Object

	UserID
0	1
1	2
2	3

Integer Series
Integers are whole numbers

	Name
0	Mike
1	Sam
2	Joe

Object Series
Text is called an object in Pandas

	Height
0	6.4
1	5.6
2	5.1

Float Series
Floating point numbers have decimals

Machine Learning Arrays

NumPy

The array that Pandas dataframes are built out of

TensorFlow

A Tensor in TensorFlow is a multi-dimensional array

MXNet

An ndarray in MXNet is almost identical to a NumPy array

One-dimensional Array

	0	1	2	3
0	1	Braund, Owen	male	52

One-dimensional Array

	0	1	2	3
0	1	0,1	male	52

One-dimensional Array

	0	1	2	3
0	1	0,1	0,2	52

Two-dimensional Array

	0	1	2	3
0	1	Braund, Owen	male	52
1	2	Cumings, John	male	38

Two-dimensional Array

	0	1	2	3
0	1	Braund, Owen	male	52
1	2	1,1	male	38

Three-dimensional Array

The Titanic Data Set

PassengerID	Name	Sex	Age	SibSp	Parch
1	Braund, Owen	male	22	1	0
2	Cumings, John	female	38	1	0

The Titanic Data Set

Survived	Fare	Cabin	Embarked	Pclass
0	7.25	NAN	S	3
1	71.28	C85	C	1

Methods and Functions

isnull()

dropna()

fillna()

duplicated()

drop_duplicates()

set_index()

Data Cleansing Steps

Remove the unnecessary columns. This will make it easier to focus on the core attributes

Identify and remove duplicates. Models like completed data set without empty or null values

Fix missing data, update incorrect data and correct format issues

Remove Unneeded Columns

PassengerID	Name	Sex	Age	SibSp	Parch
1	Braund, Owen	male	22	1	0
2	Cumings, John	female	38	1	0

Remove Unneeded Columns

Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Pclass
male	22	1	0	A/5 21171	7.25	NAN	S	3
female	38	1	0	PC 17599	71.28	C85	C	1

Remove Unneeded Columns

Sex	Age	SibSp	Parch	Fare	Cabin	Embarked	Pclass
male	22	1	0	7.25	NAN	S	3
female	38	1	0	71.28	C85	C	1

Remove Unneeded Columns

Sex	Age	SibSp	Parch	Fare	Cabin	Pclass
male	22	1	0	7.25	NAN	3
female	38	1	0	71.28	C85	1

Remove Unneeded Columns

Sex	Age	SibSp	Parch	Fare	Pclass
male	22	1	0	7.25	3
female	38	1	0	71.28	1

Remove Unneeded Columns

Sex	Age	SibSp	Parch	Fare	Pclass
1	22	1	0	7.25	3
0	38	1	0	71.28	1

Demo

- Import libraries
- Find and fill missing values
- Find and remove duplicates
- Convert sex column to integers
- Peruse model ready dataset

Summary

A series object is like a column

The array is the core data object

Importance of subject matter knowledge

The core data wrangling functions

Data cleansing guide

Completed wrangling the dataset