Course Overview

[Autogenerated] Hello. My name is Mike Weston. Welcome to my course. Data wrangling with pandas from machine learning engineers, artificial neural networks getting all the attention. One of the most overlooked aspects of machine learning is the data. Regardless of the algorithm type, almost all machine learning models need well formatted structure data to perform optimally. It's the job of the machine. They're engineer to wrangle the data into a model state. Data wrangling is one of those difficulty, time consuming parts of machine learning in the real world. Data is dirty and machine learning models or temperamental. These models only want highly structured, well cleanse data in this course provide you with the foundation you need to rankle those unruly data sets. The course will introduce you to apply data wrangling. You'll learn how developers take real world data sets and wrangle them too. Highly structured numerical entities machine learning models need the court library used by machine learning engineers to wrangle their data, and python is called pandas. You learn how to manipulate tabular data in an array. They raise the core data object in machine learning. Once the date has been properly wrangled, you'll build a highly accurate model that will predict a person's survivability if they're aboard the Titanic at the time of the sinking python has become the gold standard and apply machine learning in the library called pandas. The preferred tool utilized by developers to massage their data into a welcoming state by the end of the course will be familiar with the basics of data wrangling and the process. Machine learning engineers used to create will cleanse model ready data sets. I hope you'll join me on this journey to learn more about data wrangling with Python Plural site.

Getting Started in Data Wrangling

Introduction

[Autogenerated] Hello. My name is Mike Weston. Welcome to my core state of wrangling with pandas for machine learning engineers and this course we're gonna cover the basics of data wrangling. Machine Leonard is one of the most in demand careers in the world and will be for a long time to come. A large part of rural machine learning is data wrangling. This first module will provide some basic information about data wrangling. This includes an explanation of the two court types machine learning, supervised and unsupervised learning Machine learning is very process oriented. This first module will cover that process and explain how data wrangling fits into the larger picture Machine learning models don't like poorly structured data. The cleaner, the data, Oftentimes the better the model's performance. By the end of this module. You understand why data wrangling is so important to machine learning? We're also gonna cover the skills you need for the course and the skills you do not. Finally, we're gonna work through a simple example in python and use pandas to remove her first attributes from her data set

Why Take This Course?

[Autogenerated] machine learning is one of the most in demand careers in the world and data wrangling one of the most in demand skills within the career. Data wrangling isn't easy, and it's a time consuming endeavor. However, it's a skill all apply machine learning engineers will have to master in order secure top level position within the space. Much of apply machine learning is data wrangling. Recent surveys by car going crowd Flower found that machine learning engineers and data scientists spent up to 80% of the time wrangling data. Regardless of the trooper centers of time spent massaging data machine them. Engineers will be spending a lot of their daily routine molding their data into a model state, all supervised machine learning models in just some kind of numerical data. If that data is incomplete, the models performance will likely suffer the machine they're engineers. Goal is to build a model with a high degree of accuracy. This won't be an easy task without a welcoming state A set this course is about learning rural skills. You can immediately apply to your data sets before building your models. The cleaner. The data often times the better of the models performance

Overview of the Course

[Autogenerated] Let's take a few minutes now to go of the contents of the course. The first thing we're gonna cover is the Pandas data frame. The data frame is somewhere in concept to an Excel spreadsheet or a cable in relational database. And it's the court object in pandas we're gonna be working with and the course we're gonna clean the famous Titanic data set. The Sinking of the Titanic is one of the most infamous your Brexit history. Although there was some of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, Children and the upper class. Once our data set has been properly cleansed, we're gonna build a model from it that will help us predict a person's survivability if they were aboard at the time of the sinking. Additionally, we're gonna cover manipulating our data and pandas using a variety of functions. While our data set is structured in a tabular fashion, it does need some cleansing. We're gonna cover several techniques you'll be able to use on just about any data set you encounter for wrangling your real world data at the end of the course, we'll have a short review on what you've learned and where to go from here on your journey. To learn more about, apply data wrangling before we move ahead. Let's cover some skills you'll need for the course and some skills you will not need. You do not need to have any prior experience in machine learning. I Basic understanding would be beneficial but just not required. Well, this course will cover the machine learning process. The focus will be on one library used in data wrangling. For this course, you won't need a deep understanding of python or what the core data science libraries are. The course will introduce you to pandas, the most popular library for data wrangling and python. You won't need any advanced math skills. One of the myths surrounding machine learning engineering is that the career requires a lot of advanced mathematics and statistics. This simply isn't the case. If you made it through high school geometry and algebra, you'll be just fine. It's important to keep in mind machine learning engineers don't author any of these models. They use ones that already exist. While you don't need advanced python development skills, a basic understanding of how Python works will help you get the most out of this course. If you install python locally and know the basics of notebook navigation, then you're ready for the course and machine learning. We don't call them rows and columns. We call them attributes and observations. Regardless of the nomenclature, you'll need to have a fundamental understanding of how an Excel spreadsheet works or a table in a relational database. Well, it's true that advanced math and stats aren't required. Basic math and stats are required. For example, we're gonna be loading data into a table like structure called Honore, so an understanding of Major sees and a raise would be very beneficial.

Supervised vs. Unsupervised Learning

[Autogenerated] machine learning is divided into two separate types of tasks. Those tests were called supervised learning and unsupervised learning. The first type is called Sacrifice Machine Money and Supervised Machine Learning. We feed the machine learning model of data set. We have the answer to that nicely Cleanse data set will be fed to our model for processing, and eventually the model will make a prediction based on the patterns that finds in the attributes of that data set in our sample, we have the first five observations of the Titanic data set. This is called a subset of her data. All the observations there called the data set. Most of us are familiar with rows and columns, however, and machine learning, Rose recalled. Observations and columns were caught attributes on our data set. We have a column or an attribute that has the answer we're looking for and every row in our data set. We call this value the target variable, and our data set our target variables binary, meaning there are only two possible outcomes. A one means the passenger survived the sinking and is he remains a passenger. Did not. We asked the algorithm to learn patterns in that data set based on all the other columns excluding the target variable. In our data set, we expect the algorithm to be able to make accurate predictions in the data it's never seen. After we friended on our data set in the real world, 90% of all Marlon is done using supervise machine learning. The second type of machine learning is unsupervised and unsupervised machine learning. There is no existing data unsupervised Machine learning is a more complex process when people talk about computers teaching themselves to learn rather than us having to teach them there, often alluding to the process of unsupervised machine money unsupervised learning there was no training data set in. The outcomes are unknown. The model goes into the problem blindly and our example. The model looks at the images of humans and rabbits and is eventually able to classify them into two distinct groups. The model learns on its own and is eventually able to classify what images of rabbits and what images air humans and our example. We only have two images, however, and a real unsupervised model. The algorithm would have hundreds, if not thousands, of images, incredible as it seems unsupervised. Machine learning is the ability to solve complex problems using just the input data and the binary on off mechanisms that all computer systems were built on. There is no reference state at all.

The Machine Learning Process

[Autogenerated] machine. There is very process oriented machine. Our engineers and data scientists follow the same four steps every time they start the process of creating their protective model. Most roller machine learning is supervised. Therefore, the first steps to source the data that it comes from relational databases. And it's the number one reason the number one required skill for a machine there. Engineer is SQL. The next step is data cleansing. This is often the most time consuming part of the process, and it's a step or data wrangling takes place, and this course we're gonna spend most of our time at this step in the machine learning process. The third step in the process is model building. This is where we building tuna models. These models are complicated mathematical equations that look for patterns in our data, then make predictions based on those patterns. The final step in the process is prediction. We've built a great model, and now we're ready to feed it. Data it's never seen. The end goal of the entire process is to have a model make accurate predictions on fresh data. Fresh data is data the model has never seen before. Now you might be thinking. Where does all this data come from? Well can come from many sources. There are typically three sources of rural data. Relational databases have been around for decades. Corporations have amassed tons of structure data and every facet of the business. Much of this data remains on touch by analysis. Hello for the \_\_\_\_\_\_\_\_\_ of machine learning is changing that currently most real world machine models or source by relational databases machine learning models like clean, tabular data sets an easy way to share these data sets without the overhead of relational database is a CSP document. The word CSP is an acronym that stands for comma separated value. These documents are most prevalent with new learners because many online document repositories and tutorials you see US three documents. While a lot of modeling is currently done with relational data, traditional data is only the tip of the iceberg. 90% of all data collected globally is unstructured, or what we often referred to as big data. While this is the most prevalent data, it's also the most difficult to work with a massage in tow. Ahmad Herbal State

Removing an Attribute Demo

[Autogenerated] Let's complete a quick demo in Python 3.6, the I D for the course will be a Jupiter no book. The first step will be to import the Pandas library into a data frame and massage the data. After that, you're going to do some actual wrangling by removing an unimportant attributes. Finally, you'll peruse the results of your wrangled data, set open a Jupiter notebook and give it a name. Let's load the first line of code. The first line of code loads the Pandas library and creates an alias called PD that could be used instead of having to type out the name of the library every time the code is called an alias will allow you to type less code. Once the code is run successfully, a new cell will appear. When that cell appears, Load. Another line of code apparent is data. Frame is a lot like an Excel spreadsheet. In this line of code, we're creating a variable called data toe. Hold our data set. Next, let's load the data set into a data frame using the read See SV function. Take note your a listen. Pandas using the alias created in the code above. Once the code in the cells run, let's see what's inside the data variable. Let's use another function called the head function. This function allows you to view the first five observations of the data set. Machine learning models need clean numerical data. What attributes jumps out at you? If your thought process is focusing on clean numerical data, write the name column. Not only is it not a miracle, it's unique within the data set. If the goal of the model is to look for numerical patterns and our data, then the unique tax column won't provide our model with anything relevant. That means it has to go in order to remove that attributes. Let's using their function. The drop function can be used to remove that column from the data set in this line of code. Let's use the data variable. The houses are data set and call the drop function on the fourth column. This is a little confusing. If the goal is to remove the fourth column, then why is there a three in the line of code and python? The first position is always a zero. That means you have to start counting from the zero position. If you start counting at 20 position, the name column is that position three in the data set. Once that sells run, let's load our next line of code. The next line of code calls the head function again. The name column has been removed from our data set, and you've just taken your first step towards wrangling the data into a more model state by removing the name attributes.

Summary

[Autogenerated] that's Marshall covered. The difference between A to court types of machine learning, supervised and unsupervised supervised learning uses highly structured data for model building. Supervised learning is the most prevalent in the implied space, and it's why data wrangling is so important to machine learning. The success of the model is more dependent on data than any other facet in the process. This is not a trivial statement. Many new to this space focus too much time on model building and are often disappointed to learn that in the real world, much of your time will be spent. Wrangling data In this module, you learned about the machine learning process. More specifically, the supervised machine learning process was covered because much for applied machinery is focused on supervised learning. Lastly, you took your first step in wrangling the data set by removing the name attributes. This step was very simple, but it is indeed data wrangling

Pandas DataFrame Basics

Introduction

[Autogenerated] Clo my quest and welcome back to data wrangling and python for machine learning engineers. The second module provides some word detailed information about the core object and pandas the data frame. This includes defining the data frame, understanding why the data frame was built on top of non player raise and why we use methods to manipulate the data. The module also covered data types data types to find how data is stored. Using the correct data type is critical to the outcome of your model. For example, you can't use mathematical methods on object data type, so story numbers correctly is very important to successful data wrangling. Additionally, this module will cover the sun, your components of the data frame. The data frame has three core components and, um, are and what they are in this module and SQL. A select statement is used to retrieve data and pandas and exercise used. This module will define and show you how indexes return data after loading, merging and preparing the data set. A familiar task is to compute group statistics or possibly pivot tables for reporting, and this module group in an aggregation will also be covered. Additionally, you'll continue wrangling the Satanic data set upon completing each module, your data set will be one more step closer to the highly structured numerical data set these models need.

DataFrame Overview

[Autogenerated] The data frame is one of the core objects and pandas pandas. Data frames all share the same properties. Old data frames are two dimensional. Raise arrays were the main data structure used in machine learning and programming terminology and a raise a collection of data where each value has an index or location associated with it. Pandas data frame sit on top of numb pyre. Raise no umpires. A Python library that could be used for scientific and numerical applications. Gordon umpires short for Numerical Python, the main data structure and numb pies in India. Rey, which is shorthand for in dimensional ray when working with known pi data in an indie array, is simply referred to as an array. It's a fixed size array and memory that contains state of the same type, such as energies or floating points. The data frame was a table much like in seek war Excel. It's similar in structure, also making it possible to use somewhere operations such as aggregations, filtering and pivoting. However, because data frames are built in python, it's possible to use Python, the program more advanced operations and manipulations than seek work cell can offer. Additionally, because these data frames said on umpire raise their very fast data frames are useful because more powerful methods are built into them and python methods are associated with objects, so the data must be in a data frame to use these methods.

Data Types

[Autogenerated] on a recent demo data was loaded into a Pandas data frame. When that happened, pandas assigned data types to each column based on what the column contained. That decision had important downstream implications on how the data was manipulated. Later, data types determine what you can and what you can store in a data frame. How they distorted the data frame affects how you can manipulate it in the outputs of the calculations as well. For example, you can't perform mathematical calculations on the strangling tax data. This might seem obvious, however. Sometimes numeric values air right into a data frame of strings. In this scenario, when you try to perform calculations on string formatted numeric data, you're going to receive an error. There are three coordinated types in a Pandas data frame. The first one is the editor, 64. The state of type stores. Whole numbers 123 and four are examples of whole numbers the next day that types have flowed 64 in the stores numbers with decimal places. Even if that decimal point is a zero, we have a column that contains both energies and floating points. Pandas will sign the entire column toe afloat data type, so the decimal point you're not lost. The next core object stores text, and it's called an object data type. In addition to storing text a string, the object data type can also contain numbers. A string might be a word, a sentence for several sentences. Naming convention of the state of type was a little strange and often confusing and numb pie in Python, Texas story. It's a strain data type. The term object does tend to confuse people, and you'll often hear people say that text and pandas distorted string data type. However, the proper term is an object and pandas vernacular.

DataFrame Anatomy

[Autogenerated] Let's take some time to learn the anatomy of a data frame. The data frame is composed of three different components. The index, the columns in the data. The first object. The index represents the sequence of values on the far left hand side of the data frame. Each individual value of an index is called a label. The second set of objects are called columns. They're a sequence of values at the very top of the data friend. Each individual value. The coms is called a column but can also be referred to as a column name or a column label. Lastly, we have our values. Everything. It's House insider data frame is the data. Let's take a look at a sample data frame. You may be thinking that the unlabeled index, our data frame, look somewhere to the column label user i d. The label user ideas. Part of the data set that was imported. This is likely Amman Atomic Lee Increasing primary key from relational database. And it's not part of the structure of her data friend. The index will rarely, if ever, have a column header. Well, sometimes here, data frames referred to us tabular data This is just another name for a rectangular data with rows and columns. It's also common to refer to columns and rows is access. Collectively, they're called axes, so arose and access and a column is another axis. The word access appears as a parameter in many data frame methods. Pandas allows you to choose the direction on how the method will work with the parameter.

DataFrame Navigation

[Autogenerated] Let's take a look now a data frame navigation, the process of retrieving Onley, the rows and columns we need Independence Data frame was called indexing. If you're familiar with relational databases, then this is no different things. In a select statement on the table, the group of objects used to slice our data frames are called indexers. There are two core indexers. The two most commonly used ones are dot low and dot I look. The Yellow Sea is short for location. The first and next was called alone, and it stands for energy location. The energy is used to select rows and columns by number in the order they appear in the data friend. Remembering that I look stands for enters the location will help you remember the difference between the two the dot lok index or select state and different way than just using the index. It could select subsets of rows or columns. Most importantly, it only selects data by the label off the rose and the columns

Grouping

[Autogenerated] categorizing a dentist sad and applying a function to each group. Whether an aggregation or a transformation is often a critical aspect of a data analysis. Workflow grouping lets you slice up the rows of that airframe into well groups that have the same values and one or more categorical variables. These are useful because you can easily calculate statistics for each group. An AK agree the results into a new data frame. One of the reasons of the popularity of relational databases in SQL, which stands for a structured query language, is the ease with which data can be joined, filter transformed and aggregated. However, query languages like SQL are rather limited in the kinds of group operations that can be performed. Additionally, larger data sets often take a long time and a large amount of resource to complete. Fortunately, Panis provides all the functionality that sequel does, but returns the results much faster. Using Umpire A's, you can perform much more complex grouped operations by utilizing any function that accepts a Panoz object or a numb pyre. A. This makes grouping and aggregated and pandas much more robust than any vendor flavor of SQL and the data frame and a series all the most common aggregating attributes are accepted

Navigation Demo

[Autogenerated] like many things in machine learning, data wrangling is a process oriented endeavor. The first test goes to import the libraries needed for the project. Secondly, because this is pan digitally, the data set to work with Mr Mo the Titanic data set will be used. Once the data said isn't a variable, you'll use indexers to view various parts of it. Additionally, feel certain group the data set. Finally, you'll rang. One other attributes of the data set. Let's hope in a Jupiter notebook and give it a name. Next, let's execute one line of code at a time. I'm using the shift. Enter key shortcut on my Wonder's keyboard to execute the cell on the first line of code. The Pandas library is being imported. Once that sells be run, it's time to important data set and store it in a variable. And our sample the variable holding our data set is called data, and the next line of code the head function is being called. This will allow us to view the first five observations over data set. Skipping a few rows. Let's navigate to the cell with I look in it in this line of code I look is being run in the first access in our data set. Recall that zero is the first roller access in the data set. Let's stay with I Look, since I look is being used, you need 2 ft, energies and his parameters. This time, two images were being fed and I look the first one is a zero, and the second one is innate. This statement will return all the roads between axes, positions zero and eight. Don't forget that the access will start at zero. All right, now, let's move on to Lok In this line of code of labels being passed, the label being passed in this example is the name column. This line of code prints out all the observations of the name label. The next line of code is somewhere to the one we just executed. However, another label has been added. This label is the Age column. This line of code will return all the observations in the data set for the name and age attributes. On our next line of code. A condition is being added. That condition is equality and python to equal signs is equality, and one equal sign is assignment. This line of code is returning the name of the passenger we went past into it. Take note that the match must be exact if we leave a few letters off the name. Attribute. Python won't return the observation. In this module we learned about grouping and aggregating. Let's apply some commonly use group by functions on her data set in the next line of code, a group by is being applied to the \_\_\_ and the survived attributes using the passenger i. D column. As the item being counted, the output tells you males didn't fare so well, surviving the sinking and the next line of code, the group by function is again being used here. The group by returns the average of the mean age of the passenger in each class. Now the navigation has been covered. Let's continue wrangling the data set, and the first module name column was removed. Let's execute the drop function on the name attributes. This is the same step that was completed in the demo from Matewan. Now let's actually cute. A similar line of code against the Passenger I D column, which is at Position zero in the Axis. The passenger I D column was most likely a monotone tickly increasing key from a relational database. This has no value tour data set, so it has to go, the data says, looking much cleaner. And we've only removed two attributes from it. The two attributes that were removed for the name, attributes and the passenger attributes.

Summary

[Autogenerated] in this model, the data frame took center stage. The start of the module began detailing various core components of the data friend. The core object of machine learning for storing data is in Iraq. The data frame was a table like object that's just on top of an array. Data frames are built on top of numb pyre. Raise. This architectural decision ensures that most operations performed in a data frame are very fast. Some \_\_\_\_\_\_ tables in a relational database Data frames used data types to store various kinds of data. The three core data types are enter 64 Flowed 64 the object data type frames are composed of indexes, columns and rows. The index will often be in America, and it will be visible in the left hand side of the data frank. But navigating a data frame, you'll use indexers, the two core indexers, or I Look, which uses the index by number and local, which uses an Objects labels. Lastly, grouping aggregating in pandas are much faster, thanks to the gnome pyre, a

Pandas Data Structures

Introduction

[Autogenerated] Hello my quest and welcome back to data wrangling and python for machine learning engineers, This third module provides a more detailed information about the other core object in Panda's, the Siri's a panda Syriza's a one dimensional ray of index data. The array is the main data structure in all of machine learning. In this module, array and other similar structures will be covered. For example, Google has a framework called tensorflow. A tensor is nothing more than a multidimensional ray. In this lesson, the importance of domain knowledge will be discussed. Thus far, you've been able to pick some easy attributes to remove. Removing the name attributes in an imported primary heat didn't take a lot of real data. Set knowledge in order. Wrangle the data set much further. You'll need more detailed knowledge about the data set. Python uses a lot of methods and functions, and this monitor the core functions used in data wrangling will be covered. For example, almost all data sets have missing values. Handling missing values correctly is credible to improve your model performance. Additionally, the module will cover a road world, a bridge guide to data wrangling. This will include a step by step process of wrangling. Reward data. Finally, you'll wrangle the rest of the data set upon completion of this module, the data set will be model ready. That means the entire data set will be composed of numbers.

The Series Object

[Autogenerated] a serious represents a one dimensional labeled index array based on the non pie India Rey. Well, that's an extremely technical definition. At first to stick of a syriza's a column in a table with the index attached to it, the index is on left hand side of the Siri's. The index has no column header and begins with zero. Beside the index's air data. Recall that there are three core data types used in the data frame and the Siri's. The first is editors, and their whole numbers. Second are floating points, and they have decimals. And lastly is the object data type in its stores. Texting string data like an array A Siri's can hold zero or more values of a single data type. A Siri's can be created and initialized by passing either a scaler value in India Rey, a python list or a python dictionary.

Vendors and Arrays

[Autogenerated] the array is one of the most fundamental data objects within all of machine learning an array as a collection of data or values where each value is an index or location associated with it, regardless of the platform you use for your machine when he needs. The concept of an array is used throughout all of machine learning pandas. Data frames are built on the Empire A's when building models locally on your laptop or desktop. Nope, I will be The array used most often tensor flows Google's computational framework, and the key object is a tensor, which is a multidimensional ray. If you're building models on Google's cloud platform off to refer to his G C P, then the tensor will be the array of choice mxnet and glue on or to frameworks from Amazon. And there are a is almost identical to the numb pyrite. There are rays called in India, Rey the rays, the core object using supervised machine learning to feed data into the models. Regardless of where you bill your models, the array, our structure, somewhere to the array, will be used to house that data

Array Types

[Autogenerated] arrays have different dimensions. The most basic arrays of one dimensional ray. The data points moving only one direction, and our example. We have a one dimensional ray with four data points and any array. We use index positions to locater data if you need to locate a data point associated with passenger own. Braun, you would say this data point is located a 01 within our array. If you wanted to locate the ages, Mr Braun U U 02 to locate that data point within the array. The next type of Ray is a two dimensional ray. The data points in this array move in two directions. A set of data points moves horizontally, and another setting was vertically and our sample. If you wanted to locate the passenger with the name John Cummings, you would say he's a data 0.11 within the array. Lastly, we have a three dimensional ray. Think of a three dimensional ray as a Rubik's cube were. Each smaller cube represents a different number but moves throughout. Space is a single cube. The data points now move in three dimensions. One set of data points moves horizontally one vertically and wanted to a space that we can only view when visually represented a three dimensional object like a Rubik's Cube, each dimension with different data points on them, these three arrays, or the foundation for all data modelling and supervised learning.

The Titanic Dataset

[Autogenerated] thus far, you were able to remove certain attributes from the data set without a granular understanding of the data set. However, before moving forward and more thorough examination of the data set is in order. While names and database keys are easily removed. Data, wrangling of balls, understanding the data, your modeling and data wrangling there's no substitute for domain knowledge. Domain knowledge is knowledge about the data. Your wrangling fortunately, the Titanic, Fader said, has been hand labelled, and that means someone has gone through the data set and massage the data for you. Unfortunately, in the applied space, this will be up to you. While many of the comes are self explanatory, a few or not let's cover each of them to be thorough. The name columns, the passenger's name. This column has already been removed from the data set based on analysis from the first module. The column. Age is the age of the passenger at the time for sinking. The \_\_\_ column defines whether the passenger was male or female. The column seven p stands for someone's of passengers aboard. Did the passenger have a brother sister aboard the column? Parch means what were the total number of parents and Children aboard fair is how much the passenger pay for the ticket. The column P class stands for passenger class. There are three classes aboard the Titanic. 1st, 2nd, 3rd, first class passengers were on the upper level. Second class passengers were on the middle and third class passengers on the bottom level, the cabin columns, the room they were in and lastly embarked is a location where the passengers boarded the boat. Now you know what all the columns and attributes mean. Start thinking about what attributes you believe will contribute to the passenger survivability they to set. Familiarity is necessary in order to determine what columns will be removed from the data set. This is one of the most important first steps in the data cleansing process.

Abridged Data Cleansing Steps

[Autogenerated] While there is no data wrangling playbook, these steps are a good starting point to begin the data cleansing process. The first step is to remove any unneeded attributes from your data set. At this juncture, the question you'll need to ask of your data is what attributes can be easily removed that you know for certain, are not needed. For example, when you import data from relational database, you'll often import the primary key. Since most relational database keys or surrogate keys meaning the key has no business logic, you'll need to remove them. The second step in the process is to remove duplicate data or empty data. Recall that relational databases account for much of the data in machine learning. Relational data is often dirty, filled with holes and duplicates. You'll need to remove all those duplicate values. Once those dupes are gone, you'll need to fill those empty values in the data frame. Models don't like missing values. It's a much better practice to fill those empty value contrived or artificial data than it is to leave them blank. And the third step you want ensure that all the data is formatted properly. All punctuation will have to be cleansed. Additionally, you want to update any incorrect data the final data set will need to be in America, and we'll have all the values in the data frame filled in.

Methods and Functions

[Autogenerated] python relies heavily on methods and functions. While there are hundreds of functions and python, you'll use a few of them, often to wrangle data. Parentis offers two functions to test for missing data. There's no not know these air simple functions that return a bowling in value, indicating whether the past and are given value is, in fact missing data. You have two options for handling these values. You can replace them or you can drop them and pandas missing values show up his nan values. We call that NAN stands for not a number in order to drop all the nan values from a data frame simply called the drop in a function. Be careful, though Calling drop in a on the entire data frame will drop all occurrences of nan values in the entire data frame dropping. Only one column could be done by passing the value parameter instead of the method. Most of the time, you won't want to drop all those nan values instead of dropping all those values and other option is to call fill in a fill in a will change those values to the value you specify. The data wrangling process involves finding duplicate data and removing those duplicates in order to find duplicates in a series called The Duplicated Function. Once you found those duplicates, you can remove them by calling drop duplicates. Lastly, you often need to alter the column used in an index set. Index will allow you to change the columns, columns you want to be The index column.

Final Wrangle Demo

[Autogenerated] and this demo you'll start by important pandas and reading the data set into a data frame. These initial steps almost always be the same. Missing values are a hindrance to most models. In this demonstration, you'll learn how to find those missing values and then replace them. Relational data will often have duplicate values. Well, you can move them before exporting your data. Keep in mind that pandas data frames are much faster than most relational database systems, and this demonstration, finding and removing duplicate values will be covered. Recall The final data set will need to be numerical. Therefore, you to convert the \_\_\_ column to imagers. Lastly, you'll proves the final data said and clean anything you might have missed during the wrangling process before the final Rangel. Let's cover removing duplicates. Since there are no dupes in the Titanic data set, we're gonna have to create a contrived data set to work with On our first line of code. Pandas is being imported. Next, we're gonna create a data set filled with contrived data. When our second line of code is executed, the data set is created and the contents are printed out. Take note. There are three my quests. Well, it appears there, too, to book of values. There is really only one age value in one of those observations is 111. That's most likely an error in the data. However, it's not a duplicate value for our purposes. If this were live data, you'd have to have someone do some research to find out more information about that customer. In order to remove the dupes from the data frame, we simply called drop duplicates. After that's executed, the data from has no more duplicate values. Let's navigate back to the Titanic data set and wrangle it from the very beginning. In the first two lines of code, pandas is being imported and the data set is loaded into a data frame. Next, the head function is being called in the 1st 10 rows. From an attribute perspective, this is the entirety of the data set in the next line of code on Lee. The column selected will be returned. It's already been established that the name and passenger I D columns need to be removed. Let's return all rose, excluding the name and passenger I d columns. In this module, you learn that the tickets were randomly assigned to the passengers. Therefore, let's remove the ticket column. Embarked column is the location where the passengers boarded the boat. This is also a random occurrence, and that means the Embark column can be rang without next up is the cabin attributes. Attributes with lots of nan values tend to attract your attention in the cabin. Attribute. It appears there are lots of Dan values. Let's call the info function on all the attributes to see how many nan values the column has viewing the results you can see the cabin attributes on. Lee has 204 values, and that means the rest will be nans. If this weren't important attribute to the data set, then we would have to stop our analysis on this data set. Since the attribute doesn't provide us with any real value, it can be removed. The Fair column is the amount that passenger pay for the ticket. If the passenger class active you didn't exist, then the color might have more importance. Fortunately, the passenger class tells us where the passengers were located on the ship, and that means we could remove the fair column in this line of code, the head function was called on. The 1st 15 rows were observations. The Age column is one of our most important attributes. Women and Children first was enforced. However, The Age column does have a few nan values. The results only show one value, but one is too many. There are two options here. Removal the nan values or replace them. For this wrangle. Let's replace them. A common approach to filling Nan Values in America column is to replace them with the mean value of that attributes. In order to do that, the filling they will be called on the Age column with the mean parameter. This will replace all the nan values in that series with the mean age that mean ages 29. At this juncture, the data set looks much better and much cleaner. The one attribute that still jumps out is the \_\_\_ column. It's still text. Let's fix that by map emails to zero and females to one. In the data set, you're done. The data set is numeric, and you've removed all the unimportant attributes. This data set is now model ready. Finally, let's write out our Quinn's data, said Toe a CSP file. This is our final cleanse data set ready for the modelling process

Summary

[Autogenerated] and this marshal, the serious objects was covered. A series object is a column in a table with an index attached to it. The watch also covered the Iraq. The main day destruction, all supervised machine learning. Regardless of the platform able, your models on your data will need to be in an array or an array like object. Wrangler involves two main knowledge. In the real world, you'll have SMEs or subject matter experts you'll rely on to understand the data wrangling your data set requires you have intimate knowledge of it. The module also covered several functions you'll use often and pandas. There is no science behind filling empty or nan values. You'll have to test various models on a variety of data sets. Most of the models I've built used to mean value of the attributes. Once you began cleansing roll data, you'll develop your own routine process to get you started. This model provided you with some basic steps on how machine learning engineers approached data wrangling. Lastly, the demo walked you through a simple Rangel of the Titanic data set. Data wrangling is one of the most important skills in machine learning engineer will need to possess in order to excel in the applied space

Modeling the Cleansed Data

Introduction

[Autogenerated] Hello on my quest and welcome back to Data Rankin and Python from machine learning engineers. This fourth modules about modeling are cleanse data. Set machine learning engineers, often affectionate, referred to this is the fun part of the job building and tweaking models in order to get the best predictive capability is a rewarding process. In this model, the to court types of machine learning models will be covered. There are only two, and you'll know what they are in this module. Many machine learning models fall into four broad categories. For example, you learn the difference between classification and regression. Two categories of models that are often used in the applied space and this module, the course steps of building a predictive model in cycler will be covered. Model building like data wrangling is a process or any endeavor. And cycler has become the gold standard for building traditional machine learning models and python model selection for those new to machine learning can be very difficult. How do you know what model to choose for what project A task. This module will cover a general guide to model selection. Lastly, you're gonna be building several models using the data set that was cleansed throughout the course. Your data wrangling efforts will be rewarded with a successful real world model.

The AI Hierarchy

[Autogenerated] artificial intelligence makes it possible for machines to learn from experience, adjust new inputs and perform humanlike tasks. Artificial intelligence is the top level container that holds all of machine learning. Most examples that you hear about today from chess playing computers to self driving cars rely heavily on deep learning in natural language processing. Using these technologies, computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data machine. Learning can be defined as a set of algorithms that parse data learn from them and then apply what they've learned to make intelligent decisions. These models are often referred to as traditional models and python. These models could be built using a library called Cycle Learn and this module you'll use cycle earned a craft your models. A neural network is computer system designed to work by classifying information in the same way your brain does. It could be taught to recognize, for example, images and classified them according to elements they contain. Artificial neural networks leveraged the machine learning umbrella. You might be thinking, Why would I use a traditional model over an artificial neural network? Artificial known looks better. Well, it is true that deep learning models will often outperform their counterparts. The cost is very high. Deep learning models require a lot of data that are often very resource intensive. The model you'll be building can be constructed on a laptop with very little data and used in the real world.

The Four Machine Learning Categories

[Autogenerated] early in the coercion, or that their two broad categories of machine learning, supervised and unsupervised learning under those two broad types. For model categories, let's cover the four broad categories or techniques that fall under supervised and unsupervised lining. The first two are supervised learning tasks, and the latter two are unsupervised learning tasks, possibly the most common type of applied machinery. Misclassification, a classification model attempt to draw some conclusions from observed values. Given one of more inputs a classification model, we'll try to predict the value of one or more outcomes outcomes from labels that could be applied to a data set, for example, and this course you're building a binary classification model. There are only two possible outputs in the model, hence the term binary classification. But one means the person survived and is he remains. They did not. The next model also falls into the supervisor an umbrella, and it's linear regression. Simple linear regression is useful for finding relationships between two continuous variables. For example, predicting housing prices or stock prices are examples of linear regression. The idea is determine align the best fits the data. The next technique is called clustering, and it Falls under the unsupervised learning umbrella cluster, has a machine learning technique that the Falls, the grouping of data points. Given a set of data points, we can use a cluster an algorithm to classify each data point to a specific group and theory. Data points that are in the same group should have similar properties or features, while data points that are in different groups should have to similar properties or features. The next technique is called association, and it's an unsupervised learning Approach Association Rule based learning as a rule based machine learning method for discovering relationships between variables and large databases, it's often referred to as market basket analysis. The goal is to find associations of items that occurred together more often than you would expect from a random sampling of possibilities. The classic example was the famous Beer and Diapers Association that has often mentioned in data mining books. The story goes, men who go to the store to buy diapers will also tend to buy beer at the same time. If you were working for the beer company who discovered this unique association, then you might suggest a beer display in the eye over the diapers are sold. Machine learning is about finding patterns and making predictions and data sets. These Penner's might not seem obvious to you and I. At first, however, they make perfect sense to the algorithm.

SciKit-learn Model Building Steps

[Autogenerated] It's like it learners. A library in Python used for building traditional machine learning models. It's important to remember the library is not used for building deep learning models. It's only used for but in traditional models when building in Yamato on cycle and there are three core steps After your data has been wrangled and you've chosen the model, there are three things you need to do with every model. The first step is to fit the model. The term fitting the model is synonymous with training the model. In order for the model to the data, the fit method is called. Now that you've trained your algorithm, it's time to make some predictions in order to do that. The predict method is called. This Does Exactly As the name implies. It will make predictions best in the patterns and finds in your data. Lastly, you need evaluate the models accuracy and this course you're building a model to predict a person's survivability if they were aboard the Titanic at the time of the sinking. Classification problems are perhaps the most common type of machine learning problem, and that means there are a lot of metrics that could be used to evaluate predictions for this problem. The most common use metric is accuracy. Classification accuracy is a number of correct predictions made is a ratio of all the predictions made, you might be thinking, What's a good number for a models? Accuracy? Well, that's a very good question if you're models close to 50% and that's no better than the flip of a coin. If you're models, accuracy is 90%. Then that's a good chance you're suffering from over fitting a condition where the model learns the data to. Well, if your model is above 75 below 90 then you're probably in a good accuracy zone.

Model Selection Process

[Autogenerated] when you're given a problem, How do you know what model to choose? That's not an easy question. Fortunately, in the applied space, some decisions were made for us. This makes the model selection process a little easier. For example, you've learned the most real world models are supervised and that Mangini data before we go any further, let's walk through a simple selection process and this course we've been given a data set to model. Therefore, it's a supervised learning problem. The next question is, how much data have you been given? If you have less than 500 samples of Rose, then you need more data. The general rule is, the more data, the better. Our data set has more than 500 samples so we can move forward. The next question you only did ask is, Are you predicting a category? And is the data labeled? If the answer is yes to predict in a category, then you have a classification problem. If the data is not labeled, then it's back to data wrangling in the real world, the person doing the labeling well off to be the machine on engineer. If your data wasn't labeled, then you'd have to go back through the data set and other survived or not survived column. If you were predicting a quantity, then you look to regression. You've already learned models predicting a category, so that means you have a classification problem. Recall that there are two possible outcomes for every sample a one or a zero, and that means our model was a binary classifier. A suggested metric for deciding between a support vector machine and stochastic Grady dissent is the amount of data involved. However, one of my first choices, regardless of the amount of data, is a support vector machine model. These models tend to do very well with binary classification problems in the applied space. Many of your models will be binary classify IRS, for example, did the customer by a product. Well, the driver survived an accident. Well, the customers choose your restaurant, additionally, and cycler. You can easily switch between models, often by altering only one or two lines of code. You'll start your testing by using an SBC, but then eventually decide on a decision. Tree decision. Trees also worked very well in binary classification problems. You might be thinking, How do I know what models work on binary, classify, IRS and which ones work best on natural language processing. Unfortunately, there is no concrete answers to this question. Only time and experience weight in the model selection process.

Model Building Demo

[Autogenerated] the data set has been properly cleansed, announced time to build a model from it. And this demo we're gonna important libraries we need. In addition to importing pandas, you'll be importing models from a library called Psych It. Learn the name used for the cycle or in libraries called SK Learn. Once the data set has been property clinched, you'll build a support vector machine. This model works by creating the line that separates the data points. Once that line has been found, the model will separate any input based on where it falls specific to that line after the SEC has been scored you'll. But another model called a decision tree. A Decision tree is a set of rules used to classify data into categories it looks at. The variables on the data said determines which for the most important, then comes up with the tree of decisions that best partitions the data the trees created by splitting data up by variables and then counting to see how many are in each bucket after the split. Once that's been completed, you'll choose the model that performed the best. This data set was part of a competition host on a website called Cargo because you took the time to properly cleanse your data. The final model would have scored in the top 2% in that competition. Additionally, this model is production ready many of the winning models on cargo or never used in the real world because they're over engineered. The one you build will be highly accurate and production ready. All right, let's begin coding and the first line of code you import pandas. Next, you load the data set into a data frame, and lastly, you call the head function to peruse the data set on the next line of code you import the columns will use for the model. Let's call the head function again to ensure your data looks clean. Our data stores text in the \_\_\_ column, so let's clean that up. Once that line of code is executed, it's the head function once again to view our results and our next line of code. Let's drop all the nan values. This will reduce the data set, but also give us a true representation of the data. Let's load our next line of code, the survive columns being dropped, and it's being mapped as our target variable, and the next line of code trained test split is being imported and executed against the data set. This line of code splits the data into a training set and a testing set, and the next line of code, the support vector classifier, is being imported. No, that psych it learned calls this model and SPC for support Vector classifier. It's the same model was a support vector machine just named definitely in the Psych Ward library, and the next line of code is called the Fit Method to train the model. When you call this method, you'll see the sundry parameters are printed out. These can be adjusted to improve model performance. The defaults will do just fine for our purposes and our next line of code, the predict method. It's cold. Additionally, the model is being scored the score for this model 77%. That's not a bad first run. However, we can do better. Let's try a different model. Take note that all the steps are identical except for the two lines of code, that important decision tree model and the one that allow us to determine the splits in our tree since all the code is identical, let's just walk through until we arrived. The code. It's different. Please take note that even imported the model is almost identical to importing the SPC. The only code that's changed is the name of the model. The next line of code is the important one. The men samples split specifies the minimum number of samples required to split an internal note. Tweaking this perimeter allowed the model to score much higher. As an aside, tweaking parameters and machine learning is called hyper parameter tuning. After we fit, predict and score a model, a score of 84% was achieved. This is a great score on the state of set. Additionally, it's a role model that could be immediately placed in production. Congratulations. You've just wrangled a data set and with a few lines of code in a little bit of tweaking, you were able to score higher than 90% of all the people in the world who modeled this data set

Summary

[Autogenerated] In this model, the hierarchy of artificial intelligence was covered. Artificial intelligence is the top of container. Machine learning is a type of artificial intelligence, and lastly, artificial neural networks are a type of machine. Learning. This march also covered the core four. At a high level, models could be grouped into four different categories. When you're building a supervised learning model, you'll choose between classification and regression. When you're building an answer revised learning model, you'll choose between cluster in an association and cycle, and there are three course steps to building your model. Once your data has been cleansed and your model has been chosen, they are fit, predicted evaluate. You trained the model in the data. The model makes predictions and lastly, the models evaluated often for accuracy. Model selection is not an easy process and applied machine learning most problems were supervised. This makes the decision making process a little easier. If you've been given the data set or have to create one, they will choose between classification and regression. In the demonstration, you use the data set you wrangled throughout the course to build two different models, almost a support vector classifier or next PM in moments of decision Tree. The Decision tree model was the one ultimately chosen as the winner

Course Summary

[Autogenerated] As we wrap up this course, let's spend some time recapping our data wrangling journey. We started the course started about the to court types of machine learning, which are supervised it on supervised learning. We focused on supervised learning because most of Applied Machine Ernie evolves building predictive models against existing data sets. The machine in the process was covered with an emphasis placed on data wrangling. The single most important facet ability inaccurate machine learning models is the data. The course focused on the Pandas library in Python and an array like object called a data frame. The data frame is the key object the houses, the data you'll build your models against. You've also learned that data inside a data frame is stored based on data types. There are three core data types. Each data type holds a specific type of data. The and 64 holds enter jurors. The floats 64 holds numbers with decimal places, and the object data type holds textual information. These data frames are built on top of numb pie raise, making them very fast. The other core object and pandas. The Siri's object was covered. A serious is nothing more than a column in a table with an index attached to it. The index simply provides you with the way to access the different data points within the data frame, regardless of the platform. A use for data wrangling or your model building needs theory is the court data object on off machine learning. Tensorflow mxnet grew on numb pie. All used the same object for their data. A multidimensional ray Data wrangling is impossible. Without knowing the data, you'll need to understand what every single attribute means the applied space. That means you'll need to rely on a subject matter expert to understand the data. Data Wrangling like machine learning, is very process oriented. This course provided you with a simple, step by step approach to wrangling just about any real world data set. If you're new to machine learning, then this approach will help you get started with data wrangling. The course also covered the AI hierarchy. Artificial intelligence is apparent intended to machine learning, and machine learning is apparent. Container to artificial neural networks. There are four categories of models in the machine on the space. The two you'll use most often are classification and regression. There used most often because so much of apply machine learning is supervised. And these air supervised machine learning model categories the course covered psych it learn a library in Python. It has become the gold standard for building production. Ready traditional models. Second learners, three course steps. Once you've chosen a model and the date has been cleansed, the data is fitted to the model. Predictions are made, and lastly, the model was evaluated on how well I did. Choosing a model isn't an easy task in the applied space you off to be crafting supervised machine learning models, and that means you'll choose between classifications regression once or data was properly cleanse. We built two different models using psychic learn, a support vector machine and a decision tree. After some testing and model Tooni, the Decision tree was chosen as the winner. Congratulations. You've wrangled a data set and created a model with a high degree of accuracy that is production ready