



[Start Here](#) [Blog](#) [Books](#) [About](#) [Contact](#)

Search...



Need help with Python Machine Learning? [Take the FREE Mini-Course](#)

Your First Machine Learning Project in Python Step-By-Step

by Jason Brownlee on June 10, 2016 in [Python Machine Learning](#)



Do you want to do machine learning using Python, but you're having trouble getting started?

In this post, you will complete your first machine learning project using Python.

In this step-by-step tutorial you will:

1. Download and install Python SciPy and get the most useful package for machine learning in Python.
2. Load a dataset and understand it's structure using statistical summaries and data visualization.
3. Create 6 machine learning models, pick the best and build confidence that the accuracy is reliable.

If you are a machine learning beginner and looking to finally get started using Python, this tutorial was designed for you.

Let's get started!

- **Update Jan/2017:** Updated to reflect changes to the scikit-learn API in version 0.18.
- **Updated Mar/2017:** Added links to help setup your Python environment.



Your First Machine Learning Project in Python Step-By-Step
Photo by [cosmoflash](#), some rights reserved.

How Do You Start Machine Learning in Python?

The best way to learn machine learning is by designing and completing small projects.

Python Can Be Intimidating When Getting Started

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems.

There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird's eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

Beginners Need A Small End-to-End Project

Get Your Start in Machine Learning

Books and courses are frustrating. They give you lots of recipes and snippets, but you never get to see how they all fit together.

When you are applying machine learning to your own datasets, you are working on a project.

A machine learning project may not be linear, but it has a number of well known steps:

1. Define Problem.
2. Prepare Data.
3. Evaluate Algorithms.
4. Improve Results.
5. Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely, from loading data, summarizing data, evaluating algorithms and making some predictions.

If you can do that, you have a template that you can use on dataset after dataset. You can fill in the gaps such as further data preparation and improving result tasks later, once you have more confidence.

Hello World of Machine Learning

The best small project to start with on a new tool is the classification of iris flowers (e.g. [the iris dataset](#)).

This is a good project because it is so well understood.

- Attributes are numeric so you have to figure out how to load and handle data.
- It is a classification problem, allowing you to practice with perhaps an easier type of supervised learning algorithm.
- It is a multi-class classification problem (multi-nominal) that may require some specialized handling.
- It only has 4 attributes and 150 rows, meaning it is small and easily fits into memory (and a screen or A4 page).
- All of the numeric attributes are in the same units and the same scale, not requiring any special scaling or transforms to get started.

Let's get started with your hello world machine learning project in Python.

Machine Learning in Python: Step-By-Step Tutorial (start here)

In this section, we are going to work through a small machine learning project end-to-end.

Here is an overview of what we are going to cover:

1. Installing the Python and SciPy platform.
2. Loading the dataset.
3. Summarizing the dataset.
4. Visualizing the dataset.
5. Evaluating some algorithms.
6. Making some predictions.

Take your time. Work through each step.

Try to type in the commands yourself or copy-and-paste the commands to speed things up.

If you have any questions at all, please leave a comment at the bottom of the post.

Need help with Machine Learning in Python?

Take my free 2-week email course and discover data prep, algorithms and more (with sample code).

Click to sign-up now and also get a free PDF Ebook version of the course.

Start Your FREE Mini-Course
Now!

1. Downloading, Installing and Starting Python SciPy

Get the Python and SciPy platform installed on your system if it is not already.

I do not want to cover this in great detail, because others already have. This is already pretty straightforward, especially if you are a developer. If you do need help, ask a question in the comments.

1.1 Install SciPy Libraries

This tutorial assumes Python version 2.7 or 3.5.

There are 5 key libraries that you will need to install. Below is a list of the Python SciPy libraries required for this tutorial:

- scipy
- numpy
- matplotlib
- pandas
- sklearn

There are many ways to install these libraries. My best advice is to pick one method then be consistent in installing each library.

The [scipy installation page](#) provides excellent instructions for installing the above libraries on multiple different platforms, such as Linux, mac OS X and Windows. If you have any doubts or questions, refer to this guide, it has been followed by thousands of people.

- On Mac OS X, you can use macports to install Python 2.7 and these libraries. For more information on macports, [see the homepage](#).
- On Linux you can use your package manager, such as yum on Fedora to install RPMs.

If you are on Windows or you are not confident, I would recommend installing the free version of [Anaconda](#) that includes everything you need.

Note: This tutorial assumes you have scikit-learn version 0.18 or higher installed.

Need more help? See one of these tutorials:

- [How to Setup a Python Environment for Machine Learning and Deep Learning with Anaconda](#)
- [How to Create a Linux Virtual Machine For Machine Learning Development With Python 3](#)

1.2 Start Python and Check Versions

It is a good idea to make sure your Python environment was installed successfully and is working as expected.

The script below will help you test out your environment. It imports each library required in this tutorial and prints the version.

Open a command line and start the python interpreter:

```
1 python
```

I recommend working directly in the interpreter or writing your scripts and running them on the command line rather than big editors and IDEs. Keep things simple and focus on the machine learning not the toolchain.

Type or copy and paste the following script:

```

1 # Check the versions of libraries
2
3 # Python version
4 import sys
5 print('Python: {}'.format(sys.version))
6 # scipy
7 import scipy
8 print('scipy: {}'.format(scipy.__version__))
9 # numpy
10 import numpy
11 print('numpy: {}'.format(numpy.__version__))
12 # matplotlib
13 import matplotlib
14 print('matplotlib: {}'.format(matplotlib.__version__))
15 # pandas
16 import pandas
17 print('pandas: {}'.format(pandas.__version__))
18 # scikit-learn
19 import sklearn
20 print('sklearn: {}'.format(sklearn.__version__))

```

Here is the output I get on my OS X workstation:

```

1 Python: 2.7.11 (default, Mar 1 2016, 18:40:10)
2 [GCC 4.2.1 Compatible Apple LLVM 7.0.2 (clang-700.1.81)]
3 scipy: 0.17.0
4 numpy: 1.10.4
5 matplotlib: 1.5.1
6 pandas: 0.17.1
7 sklearn: 0.18.1

```

Compare the above output to your versions.

Ideally, your versions should match or be more recent. The APIs do not change quickly, so do not be too concerned if you are a few versions behind, Everything in this tutorial will very likely still work for you.

If you get an error, stop. Now is the time to fix it.

If you cannot run the above script cleanly you will not be able to complete this tutorial.

My best advice is to Google search for your error message or post a question on [Stack Exchange](#).

2. Load The Data

We are going to use the iris flowers dataset. This dataset is famous because it is used as the “hello world” dataset in machine learning and statistics by pretty much everyone.

The dataset contains 150 observations of iris flowers. There are four columns of measurements of the flowers in centimeters. The fifth column is the species of the flower observed. All observed flowers belong to one of three species.

You can [learn more about this dataset on Wikipedia](#).

In this step we are going to load the iris data from CSV file URL.

2.1 Import libraries

First, let's import all of the modules, functions and objects we are going to use in this tutorial.

```

1 # Load libraries
2 import pandas
3 from pandas.plotting import scatter_matrix
4 import matplotlib.pyplot as plt
5 from sklearn import model_selection
6 from sklearn.metrics import classification_report
7 from sklearn.metrics import confusion_matrix
8 from sklearn.metrics import accuracy_score
9 from sklearn.linear_model import LogisticRegression
10 from sklearn.tree import DecisionTreeClassifier
11 from sklearn.neighbors import KNeighborsClassifier
12 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
13 from sklearn.naive_bayes import GaussianNB
14 from sklearn.svm import SVC

```

Everything should load without error. If you have an error, stop. You need a working SciPy environment before continuing. See the advice above about setting up your environment.

2.2 Load Dataset

We can load the data directly from the UCI Machine Learning repository.

We are using pandas to load the data. We will also use pandas next to explore the data both with descriptive statistics and data visualization.

Note that we are specifying the names of each column when loading the data. This will help later when we explore the data.

```
1 # Load dataset
2 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
3 names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
4 dataset = pandas.read_csv(url, names=names)
```

The dataset should load without incident.

If you do have network problems, you can download the [iris.data](#) file into your working directory and load it using the same method, changing URL to the local file name.

3. Summarize the Dataset

Now it is time to take a look at the data.

In this step we are going to take a look at the data a few different ways:

1. Dimensions of the dataset.
2. Peek at the data itself.
3. Statistical summary of all attributes.
4. Breakdown of the data by the class variable.

Don't worry, each look at the data is one command. These are useful commands that you can use again and again on future projects.

3.1 Dimensions of Dataset

We can get a quick idea of how many instances (rows) and how many attributes (columns) the data contains with the shape property.

```
1 # shape
2 print(dataset.shape)
```

You should see 150 instances and 5 attributes:

```
1 (150, 5)
```

3.2 Peek at the Data

It is also always a good idea to actually eyeball your data.

```
1 # head
2 print(dataset.head(20))
```

You should see the first 20 rows of the data:

	sepal-length	sepal-width	petal-length	petal-width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

3.3 Statistical Summary

Now we can take a look at a summary of each attribute.

This includes the count, mean, the min and max values as well as some percentiles.

```
1 # descriptions
2 print(dataset.describe())
```

We can see that all of the numerical values have the same scale (centimeters) and similar ranges between 0 and 8 centimeters.

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

3.4 Class Distribution

Let's now take a look at the number of instances (rows) that belong to each class. We can view this as an absolute count.

```
1 # class distribution
2 print(dataset.groupby('class').size())
```

We can see that each class has the same number of instances (50 or 33% of the dataset).

class	
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

4. Data Visualization

We now have a basic idea about the data. We need to extend that with some visualizations.

We are going to look at two types of plots:

1. Univariate plots to better understand each attribute.
2. Multivariate plots to better understand the relationships between attributes.

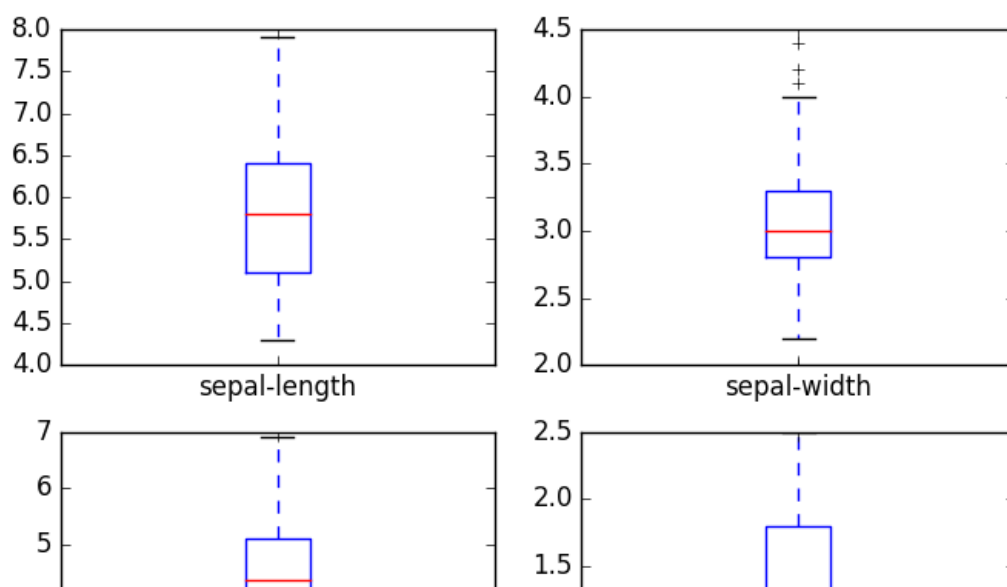
4.1 Univariate Plots

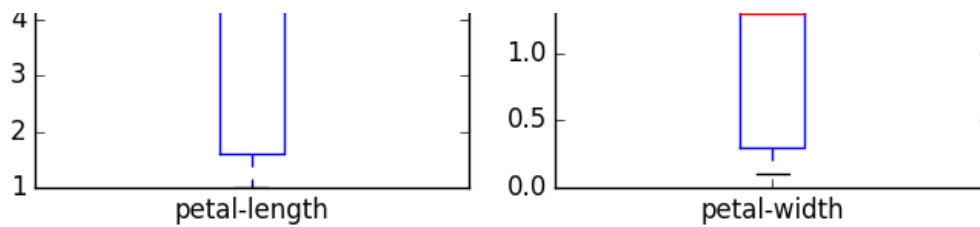
We start with some univariate plots, that is, plots of each individual variable.

Given that the input variables are numeric, we can create box and whisker plots of each.

```
1 # box and whisker plots
2 dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
3 plt.show()
```

This gives us a much clearer idea of the distribution of the input attributes:



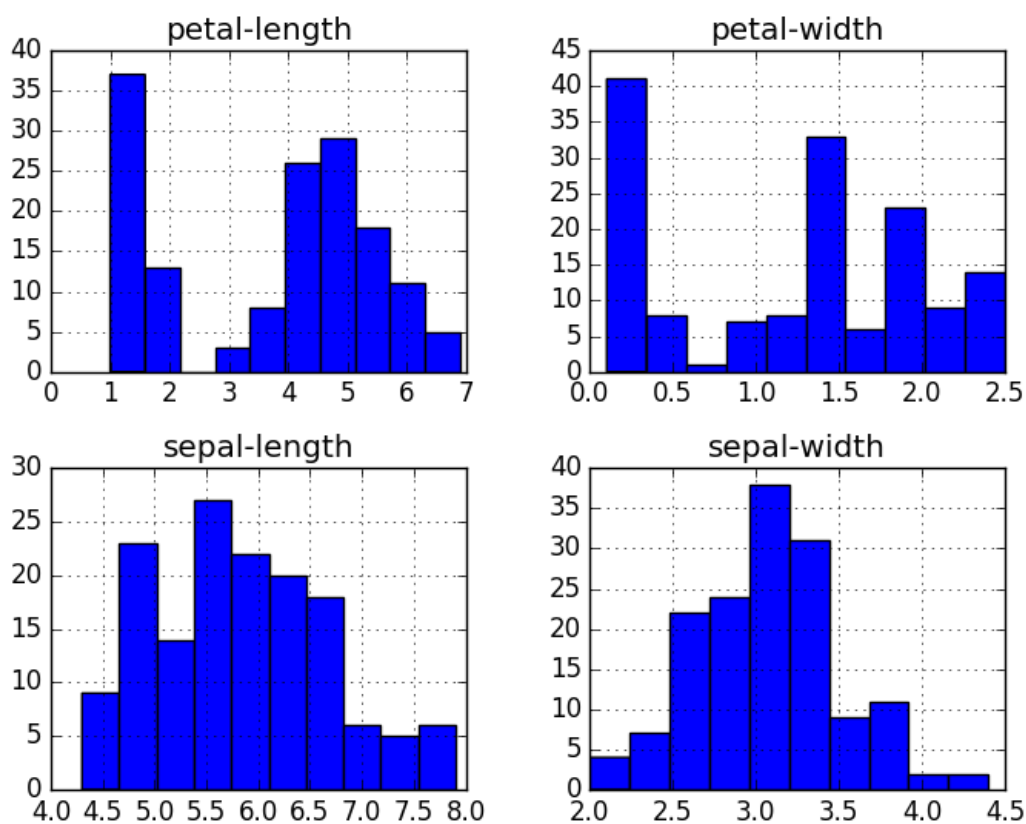


Box and Whisker Plots

We can also create a histogram of each input variable to get an idea of the distribution.

```
1 # histograms
2 dataset.hist()
3 plt.show()
```

It looks like perhaps two of the input variables have a Gaussian distribution. This is useful to note as we can use algorithms that can exploit this assumption.



Histogram Plots

4.2 Multivariate Plots

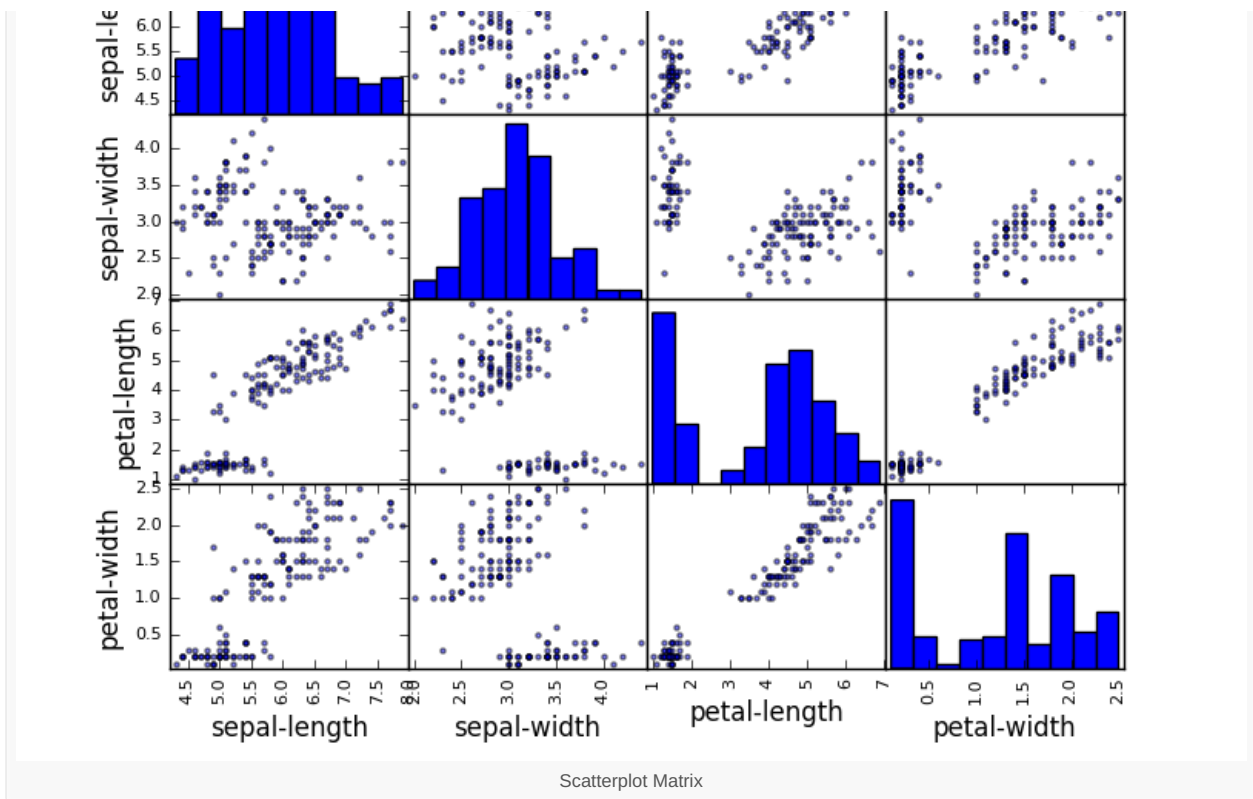
Now we can look at the interactions between the variables.

First, let's look at scatterplots of all pairs of attributes. This can be helpful to spot structured relationships between input variables.

```
1 # scatter plot matrix
2 scatter_matrix(dataset)
3 plt.show()
```

Note the diagonal grouping of some pairs of attributes. This suggests a high correlation and a predictable relationship.





5. Evaluate Some Algorithms

Now it is time to create some models of the data and estimate their accuracy on unseen data.

Here is what we are going to cover in this step:

1. Separate out a validation dataset.
2. Set-up the test harness to use 10-fold cross validation.
3. Build 5 different models to predict species from flower measurements
4. Select the best model.

5.1 Create a Validation Dataset

We need to know that the model we created is any good.

Later, we will use statistical methods to estimate the accuracy of the models that we create on unseen data. We also want a more concrete estimate of the accuracy of the best model on unseen data by evaluating it on actual unseen data.

That is, we are going to hold back some data that the algorithms will not get to see and we will use this data to get a second and independent idea of how accurate the best model might actually be.

We will split the loaded dataset into two, 80% of which we will use to train our models and 20% that we will hold back as a validation dataset.

```

1 # Split-out validation dataset
2 array = dataset.values
3 X = array[:,0:4]
4 Y = array[:,4]
5 validation_size = 0.20
6 seed = 7
7 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)

```

You now have training data in the *X_train* and *Y_train* for preparing models and a *X_validation* and *Y_validation* sets that we can use later.

5.2 Test Harness

We will use 10-fold cross validation to estimate accuracy.

This will split our dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits.

```

1 # Test options and evaluation metric
2 seed = 7

```



```
3 scoring = 'accuracy'
```

We are using the metric of 'accuracy' to evaluate models. This is a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate). We will be using the *scoring* variable when we run build and evaluate each model next.

5.3 Build Models

We don't know which algorithms would be good on this problem or what configurations to use. We get an idea from the plots that some of the classes are partially linearly separable in some dimensions, so we are expecting generally good results.

Let's evaluate 6 different algorithms:

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM).

This is a good mixture of simple linear (LR and LDA), nonlinear (KNN, CART, NB and SVM) algorithms. We reset the random number seed before each run to ensure that the evaluation of each algorithm is performed using exactly the same data splits. It ensures the results are directly comparable.

Let's build and evaluate our five models:

```
1 # Spot Check Algorithms
2 models = []
3 models.append(('LR', LogisticRegression()))
4 models.append(('LDA', LinearDiscriminantAnalysis()))
5 models.append(('KNN', KNeighborsClassifier()))
6 models.append(('CART', DecisionTreeClassifier()))
7 models.append(('NB', GaussianNB()))
8 models.append(('SVM', SVC()))
9 # evaluate each model in turn
10 results = []
11 names = []
12 for name, model in models:
13     kfold = model_selection.KFold(n_splits=10, random_state=seed)
14     cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
15     results.append(cv_results)
16     names.append(name)
17     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
18     print(msg)
```

5.4 Select Best Model

We now have 6 models and accuracy estimations for each. We need to compare the models to each other and select the most accurate.

Running the example above, we get the following raw results:

```
1 LR: 0.966667 (0.040825)
2 LDA: 0.975000 (0.038188)
3 KNN: 0.983333 (0.033333)
4 CART: 0.975000 (0.038188)
5 NB: 0.975000 (0.053359)
6 SVM: 0.981667 (0.025000)
```

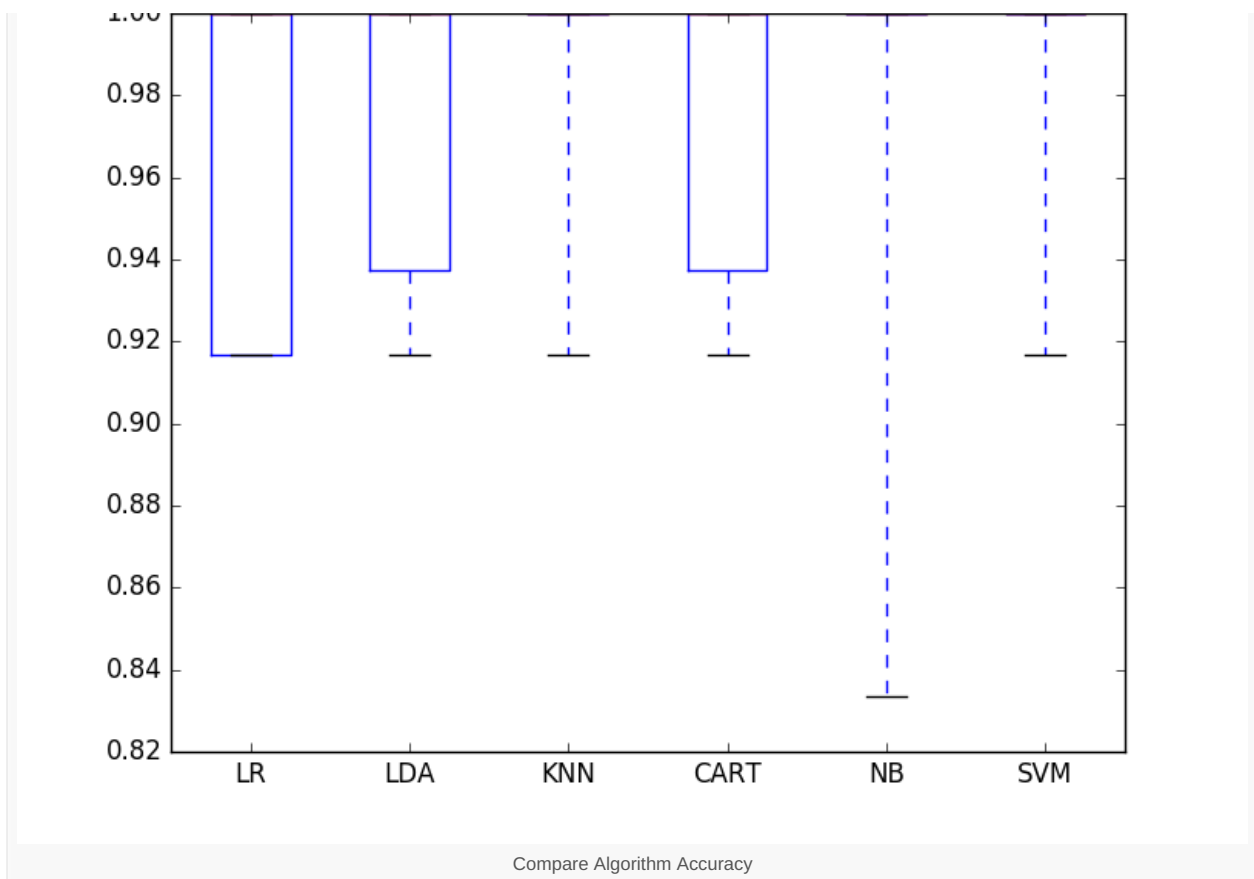
We can see that it looks like KNN has the largest estimated accuracy score.

We can also create a plot of the model evaluation results and compare the spread and the mean accuracy of each model. There is a population of accuracy measures for each algorithm because each algorithm was evaluated 10 times (10 fold cross validation).

```
1 # Compare Algorithms
2 fig = plt.figure()
3 fig.suptitle('Algorithm Comparison')
4 ax = fig.add_subplot(111)
5 plt.boxplot(results)
6 ax.set_xticklabels(names)
7 plt.show()
```

You can see that the box and whisker plots are squashed at the top of the range, with many samples achieving 100% accuracy.

Algorithm Comparison



6. Make Predictions

The KNN algorithm was the most accurate model that we tested. Now we want to get an idea of the accuracy of the model on our validation set.

This will give us an independent final check on the accuracy of the best model. It is valuable to keep a validation set just in case you made a slip during training, such as overfitting to the training set or a data leak. Both will result in an overly optimistic result.

We can run the KNN model directly on the validation set and summarize the results as a final accuracy score, a [confusion matrix](#) and a classification report.

```
1 # Make predictions on validation dataset
2 knn = KNeighborsClassifier()
3 knn.fit(X_train, Y_train)
4 predictions = knn.predict(X_validation)
5 print(accuracy_score(Y_validation, predictions))
6 print(confusion_matrix(Y_validation, predictions))
7 print(classification_report(Y_validation, predictions))
```

We can see that the accuracy is 0.9 or 90%. The confusion matrix provides an indication of the three errors made. Finally, the classification report provides a breakdown of each class by precision, recall, f1-score and support showing excellent results (granted the validation dataset was small).

```
1 0.9
2
3 [[ 7  0  0]
4  [ 0 11  1]
5  [ 0  2  9]]
6
7      precision    recall  f1-score   support
8
9  Iris-setosa       1.00      1.00      1.00         7
10 Iris-versicolor   0.85      0.92      0.88        12
11 Iris-virginica    0.90      0.82      0.86        11
12
13 avg / total       0.90      0.90      0.90        30
```

You Can Do Machine Learning in Python

Work through the tutorial above. It will take you 5-to-10 minutes, max!

You do not need to understand everything. (at least not right now) Your goal is to run through the tutorial end-to-end and get a result. You do not need to understand everything on the first pass. List down your questions as you go. Make heavy use of the

`help("FunctionName")` help syntax in Python to learn about all of the functions that you're using.

You do not need to know how the algorithms work It is important to know about the limitations and how to configure machine learning algorithms. But learning about algorithms can come later. You need to build up this algorithm knowledge slowly over a long period of time. Today, start off by getting comfortable with the platform.

You do not need to be a Python programmer The syntax of the Python language can be intuitive if you are new to it. Just like other languages, focus on function calls (e.g. `function()`) and assignments (e.g. `a = "b"`). This will get you most of the way. You are a developer, you know how to pick up the basics of a language real fast. Just get started and dive into the details later.

You do not need to be a machine learning expert You can learn about the benefits and limitations of various algorithms later, and there are plenty of posts that you can read later to brush up on the steps of a machine learning project and the importance of evaluating accuracy using cross validation.

What about other steps in a machine learning project We did not cover all of the steps in a machine learning project because this is your first project and we need to focus on the key steps. Namely, loading data, looking at the data, evaluating some algorithms and making some predictions. In later tutorials we can look at other data preparation and result improvement tasks.

Summary

In this post, you discovered step-by-step how to complete your first machine learning project in Python.

You discovered that completing a small end-to-end project from loading the data to making predictions is the best way to get familiar with a new platform.

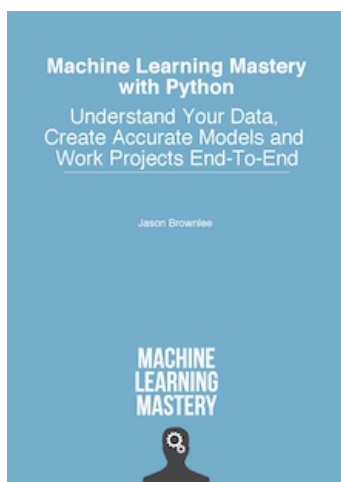
Your Next Step

Do you work through the tutorial?

1. Work through the above tutorial.
2. List any questions you have.
3. Search or research the answers.
4. Remember, you can use the `help("FunctionName")` in Python to get help on any function.

Do you have a question? Post it in the comments below.

Frustrated With Python Machine Learning?



Develop Your Own Models in Minutes

...with just a few lines of scikit-learn code

Discover how in my new Ebook:

[Machine Learning Mastery With Python](#)

Covers **self-study tutorials** and **end-to-end projects** like:
Loading data, visualization, modeling, tuning, and much more...

Finally Bring Machine Learning To Your Own Projects

Skip the Academics. Just Results.

[Click to learn more](#) .



About Jason Brownlee



Jason Brownlee, Ph.D. is a machine learning specialist who teaches developers how to get results with modern machine learning methods via hands-on tutorials.

[View all posts by Jason Brownlee](#) →

< Regression Tutorial with the Keras Deep Learning Library in Python

Save and Load Your Keras Deep Learning Models >

857 Responses to *Your First Machine Learning Project in Python Step-By-Step*



DR Venugopala Rao Manneni June 11, 2016 at 5:58 pm #

REPLY ↩

Awesome... But in your Blog please introduce SOM (Self Organizing maps) for unsupervised methods and also add printing parameters (Coefficients)code.



Jason Brownlee June 14, 2016 at 8:17 am #

REPLY ↩

I generally don't cover unsupervised methods like clustering and projection methods.

This is because I mainly focus on and teach predictive modeling (e.g. classification and regression) and I just don't find unsupervised methods that useful.



Rajesh January 21, 2018 at 5:33 pm #

REPLY ↩

Jason,
Can you elaborate what you don't find unsupervised methods useful?



Jason Brownlee January 22, 2018 at 4:42 am #

REPLY ↩

Because my focus is predictive modeling.



Hasnain July 8, 2017 at 8:55 pm #

REPLY ↩

I have installed all libraries that were in your How to Setup Python environment... blog. All went fine but when i run the starting imports code I get error at first line "ModuleNotFoundError: No module named 'pandas'". But I did install it using "pip install pandas" command. I am working on a windows machine.



Jason Brownlee July 9, 2017 at 10:53 am #

REPLY ↩

Sorry to hear that. Consider rebooting your machine?



Sheila Dawn August 9, 2017 at 5:43 am #

REPLY ↩

I had the same problem initially, because I made 2 python files.. one for loading the libraries, and another for loading the iris dataset.

Then I decided to put the two commands in one python file, it solved problem. 😊



Jason Brownlee August 9, 2017 at 6:43 am #

Yes, all commands go in the one file. Sorry for the confusion.



Dan Fiorino July 16, 2017 at 2:37 am #

REPLY ↩

Hasnain, try setting the environment variable PYTHON_PATH and PATH to include the path to the site packages of the version of python you have permission to alter

```
export PYTHONPATH="$PYTHONPATH:/path/to/Python/2.7/site-packages/"
export PATH="$PATH:/path/to/Python/2.7/site-packages/"
```

obviously replacing "/path/to" with the actual path. My system Python is in my /Users//Library folder but I'm on a Mac.

You can add the export lines to a script that runs when you open a terminal ("~/bash_profile" if you use BASH).

That might not be 100% right, but it should help you on your way.



Jason Brownlee July 16, 2017 at 8:00 am #

REPLY ↩

Thanks for posting the tip Dan, I hope it helps.



Jason Robinette September 7, 2017 at 11:16 am #

got it to work have no idea how but it worked! I am like the kid at t-ball that closes his eyes and takes a swing!



Jason Brownlee September 7, 2017 at 12:58 pm #

I'm glad to hear that!



Tanya September 30, 2017 at 11:08 am #

REPLY ↩

I am starting at square 0, and after clearing a first few hurdles, I was not even able to install the libraries at all... (as a newb), I didn't see where I even GO to import this:

```
# Load libraries
import pandas
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```



Jason Brownlee October 1, 2017 at 9:04 am #

REPLY ↩

Perhaps this step-by-step tutorial will help you set up your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



KASINATH PS December 7, 2017 at 8:16 pm #

REPLY ↩

if u r using python 3

save all the commands as a py file
then in a python shell enter

```
exec(open("[path to file with name]").read())
```

if u open shell in the same path as the saved thing
then u only need to enter the filename alone

ex:

lets say i saved it as load.py

then

```
exec(open("load.py").read())
```

this will execute all commands in the current shell



Rahul December 7, 2017 at 10:28 pm #

REPLY ↩

Hi Tanya,

This tutorial is so intuitive that I went through this tutorial with a breeze.

Install PyCharm from JetBrains available here <https://www.jetbrains.com/pycharm/download/download-thanks.html?platform=windows&code=PCC>

Install PIP (The de-facto python package manager) and then click "Terminal" in PyCharm to bring up the interactive DOS like terminal. Once you have installed PIP then there you can issue the following commands:

```
pip install numpy
```

```
pip install scipy
```

```
pip install matplotlib
```

```
pip install pandas
```

```
pip install sklearn
```

All other steps in the tutorial are valid and do not need a single line of change apart from where its mentioned

from pandas.tools.plotting import scatter_matrix , change it to

from pandas.plotting import scatter_matrix



Jason Brownlee December 8, 2017 at 5:39 am #

Thanks for the tips Rahul.



Murtaza December 17, 2017 at 11:05 am #

For a beginner i believe Anacondas Jupyter notebooks would be the best option. As they can include markdown for future reference which is essential as beginner (backpropogation :p). But again varies person to person



Jason Brownlee December 18, 2017 at 5:19 am #

I find notebooks confuse beginners more than help.

Running a Python script on the command line is so much simpler.



Jason March 1, 2018 at 4:18 pm #

Except for me, on Debian Stretch with pandas 0.19.2, I had to use

from pandas.tools.plotting import scatter_matrix



Jason Brownlee March 2, 2018 at 5:30 am #

You must update your version of Pandas



avanish March 25, 2018 at 7:11 pm #

REPLY ↩

use jupyter notebook ...there all the essential libraries are preinstalled



Jan de Lange June 20, 2016 at 10:43 pm #

REPLY ↩

Nice work Jason. Of course there is a lot more to tell about the code and the Models applied if this is intended for people starting out with ML (like me). Rather than telling which "button to press" to make work, it would be nice to know why also. I looked at a sample of your book (advanced) if you are covering the why also, but it looks like it's limited?

On this particular example, in my case SVM reached 99.2% and was thus the best Model. I gather this is because the test and training sets are drawn randomly from the data.



Jason Brownlee June 21, 2016 at 7:04 am #

REPLY ↩

This tutorial and the book are laser focused on how to use Python to complete machine learning projects.

They already assume you know how the algorithms work.

If you are looking for background on machine learning algorithms, take a look at this book:

<https://machinelearningmastery.com/master-machine-learning-algorithms/>



Alan July 26, 2017 at 10:50 pm #

REPLY ↩

Jan de Lange and Jason,

Before anything else, I truly like to thank Jason for this wonderful, concise and practical guideline on using ML for solving a predictive problem.

In terms of the example you have provided, I can confirm 'Jan de Lange' 's outcome. I've got the same accuracy result for SVM (0.991667 to be precise). I've just upgraded the Canopy version I had installed on my machine to version 2.1.3.3542 (64 bit) and your reasoning makes sense that this discrepancy could be because of its random selection of data. But this procedure could open up a new 'can of worms' as some say. since the selection of best model is on the line.

Thank you again Jason for this practical article on ML.



Jason Brownlee July 27, 2017 at 8:06 am #

REPLY ↩

Thanks Alan.

Absolutely. Machine learning algorithms are stochastic. This is a feature, not a bug. It helps us move through the landscape of possible models efficiently.

See this post:

<http://machinelearningmastery.com/randomness-in-machine-learning/>

And this post on finalizing a model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>

Does that help?



Per December 15, 2017 at 7:36 pm #

REPLY ↩

Got it working too, changing the scatter_matrix import like Rahul did.
But I also had to install tkinter first (yum install tkinter).

Very nice tutorial, Jason!



Jason Brownlee December 16, 2017 at 5:24 am #

REPLY ↩

Glad to hear it!



Nil June 25, 2016 at 12:42 am #

REPLY ↩

Awesome, I have tested the code it is impressive. But how could I use the model to predict if it is Iris-setosa or Iris-versicolor or Iris-virginica when I am given some values representing sepal-length, sepal-width, petal-length and petal-width attributes?



Jason Brownlee June 25, 2016 at 5:09 am #

REPLY ↩

Great question. You can call `model.predict()` with some new data.

For an example, see Part 6 in the above post.



JamieFox March 28, 2017 at 6:38 am #

REPLY ↩

Dear Jason Brownlee, I was thinking about the same question of Nil. To be precise I was wondering how can I know, after having seen that my model has a good fit, which values of sepal-length, sepal-width, petal-length and petal-width corresponds to Iris-setosa eccc..

For instance, if I have p predictors and two classes, how can I know which values of the predictors blend to one class or the other. Knowing the value of predictors allows me to use the model in the daily operativity. Thx



Jason Brownlee March 28, 2017 at 8:27 am #

REPLY ↩

Not knowing the statistical relationship between inputs and outputs is one of the down sides of using neural networks.



JamieFox March 29, 2017 at 7:03 am #

Hi Mr Jason Brownlee, thks for your answer. So all algorithms, such as SVM, LDA, random forest.. have this drawbacks? Can you suggest me something else?

Because logistic regression is not like this, or am I wrong?



Jason Brownlee March 29, 2017 at 9:14 am #

All algorithms have limitations and assumptions. For example, Logistic Regression makes assumptions about the distribution of variates (Gaussian) and more:

https://en.wikipedia.org/wiki/Logistic_regression

Nevertheless, we can make useful models (skillful) even when breaking assumptions or pushing past limitations.



Sujon September 6, 2016 at 8:19 am #

REPLY ↩

Dear Sir,

It seems I'm in the right place in right time! I'm doing my master thesis in machine learning from Stockholm University. Could you give me some references for laughter audio conversation to CSV file? You can send me anything on sujon2100@gmail.com. Thanks a lot and wish your very best and will keep in touch.



Sujon September 6, 2016 at 8:32 am #

REPLY ↩

Sorry I mean laughter audio to CSV conversion.



Jason Brownlee September 6, 2016 at 9:49 am #

REPLY ↩

Sorry, I have not seen any laughter audio to CSV conversion tools/techniques.



Sujon May 10, 2017 at 1:02 pm #

REPLY ↩

Hi again, do you have any publication of this article "Your First Machine Learning Project in Python Step-By-Step"? Or any citation if you know? Thanks.



Jason Brownlee May 11, 2017 at 8:28 am #

REPLY ↩

No, you can reference the blog post directly.



Roberto U September 19, 2016 at 9:17 am #

REPLY ↩

Sweet way of condensing monstrous amount of information in a one-way street. Thanks!

Just a small thing, you are creating the Kfold inside the loop in the cross validation. Then, you use the same seed to keep the comparison across predictors constant.

That works, but I think it would be better to take it out of the loop. Not only is more efficient, but it is also much immediately clearer that all predictors are using the same Kfold.

You can still justify the use of the seeds in terms of replicability; readers getting the same results on their machines.

Thanks again!



Jason Brownlee September 20, 2016 at 8:27 am #

REPLY ↩

Great suggestion, thanks Roberto.



Francisco September 20, 2016 at 2:02 am #

REPLY ↩

Hello Jaso.

Thank you so much for your help with Machine Learning and congratulations for your excellent website.

I am a beginner in ML and DeepLearning. Should I download Python 2 or Python 3?

Thank you very much.

Francisco



Jason Brownlee September 20, 2016 at 8:33 am #

REPLY ↩

I use Python 2 for all my work, but my students report that most of my examples work in Python 3 with little change.



ShawnJ October 11, 2016 at 5:24 am #

REPLY ↩

Jason,

Thank you so much for putting this together. I am been a software developer for almost two decades and am getting interested in machine learning. Found this tutorial accurate, easy to follow and very informative.



Jason Brownlee October 11, 2016 at 7:24 am #

REPLY ↩

Thanks ShawnJ, I'm glad you found it useful.



Wendy G October 14, 2016 at 5:37 am #

REPLY ↩

Jason,

Thanks for the great post! I am trying to follow this post by using my own dataset, but I keep getting this error "Unknown label type: array ([some numbers from my dataset])". So what's the problem on earth, any possible solutions?

Thanks,



Jason Brownlee October 14, 2016 at 9:08 am #

REPLY ↩

Hi Wendy,

Carefully check your data. Maybe print it on the screen and inspect it. You may have some string values that you may need to convert to numbers using data preparation.



fara October 20, 2016 at 7:15 am #

REPLY ↩

hi thanks for great tutorial, i'm also new to ML...this really helps but i was wondering what if we have non-numeric values? i have mixture of numeric and non-numeric data and obviously this only works for numeric. do you also have a tutorial for that or would you please send me a source for it? thank you



Jason Brownlee October 20, 2016 at 8:41 am #

REPLY ↩

Great question fara.

We need to convert everything to numeric. For categorical values, you can convert them to integers (label encoding) and then to new binary features (one hot encoding).



fara October 20, 2016 at 8:53 am #

REPLY ↩

after I post my comment here i saw this: "DictVectorizer" i think i can use it for converting non-numeric to numeric, right?



Jason Brownlee October 20, 2016 at 11:15 am #

REPLY ↩

I would recommend the LabelEncoder class followed by the OneHotEncoder class in scikit-learn.

I believe I have tutorials on these here:

<http://machinelearningmastery.com/data-preparation-gradient-boosting-xgboost-python/>



fara October 21, 2016 at 3:53 am #

thank you it's great



Mazhar Dootio October 23, 2016 at 9:14 pm #

REPLY ↩

Hello Jason

Thank you for publishing this great machine learning tutorial.

It is really awesome awesome awesome.....!

I test your tutorial on python-3 and it works well but what I face here is to load my data set from my local drive. I followed your give

I test your tutorial on python 3 and it works well but what I face here is to load my data set from my local drive. I followed your give instructions but couldn't be successful.

My syntax is as under:

```
import unicodedata
url = open(r'C:\Users\mazhar\Anaconda3\Lib\site-packages\sindhi2.csv', encoding='utf-8').readlines()
names = ['class', 'sno', 'gender', 'morphology', 'stem', 'fword']
dataset = pandas.read_csv(url, names=names)
```

python-3 jupyter notebook does not loads this. Kindly help me in regard.



Jason Brownlee October 24, 2016 at 7:05 am #

REPLY ↩

Hi Mazhar, thanks.

Are you able to load the file on the command line away from the notebook?

Perhaps the notebook environment is causing trouble?



Kenny October 11, 2017 at 3:43 am #

REPLY ↩

Mazhar try this:

import pandas as pd

.

file= \"namefile.csv\" #or c:/ ____/ ____/

df = pd.read_csv(file)

in Jupyter

<https://www.anaconda.com/download/>

<https://anaconda.org/anaconda/python>



Mazhar Dootio October 25, 2016 at 3:22 am #

REPLY ↩

Dear Jason

Thank you for response

I am using Python 3 with anaconda jupyter notebook

so which python version you would like to suggest me and kindly write here syntax of opening local dataset file from local drive that how can I load utf-8 dataset file from my local drive.



Jason Brownlee October 25, 2016 at 8:32 am #

REPLY ↩

Hi Mazhar, I teach using Python 2.7 with examples from the command line.

Many of my students report that the code works in Python 3 and in notebooks with little or no changes.



Kenny October 11, 2017 at 3:50 am #

REPLY ↩

try with this command:

df = pd.read_csv(file, encoding='latin-1') #if you are working with csv “,” or “;” put sep='|',



Andy October 27, 2016 at 11:59 pm #

REPLY ↩

Great tutorial but perhaps I'm missing something here. Let's assume I already know what model to use (perhaps because I know the data well... for example).

```
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
```

I then use the models to predict:

```
print(knn.predict(an array of variables of a record I want to classify))
```

Is this where the whole ML happens?

```
knn.fit(X_train, Y_train)
```

What's the difference between this and say a non ML model/algorithm? Is it that in a non ML model I have to find the coefficients/parameters myself by statistical methods?; and in the ML model the machine does that itself?

If this is the case then to me it seems that a researcher/coder did most of the work for me and wrap it in a nice function. Am I missing something? What is special here?



Jason Brownlee October 28, 2016 at 9:14 am #

REPLY ↩

Hi Andy,

Yes, your comment is generally true.

The work is in the library and choice of good libraries and training on how to use them well on your project can take you a very long way very quickly.

Stats is really about small data and understanding the domain (descriptive models). Machine learning, at least in common practice, is leaning towards automation with larger datasets and making predictions (predictive modeling) at the expense of model interpretation/understandability. Prediction performance trumps traditional goals of stats.

Because of the automation, the focus shifts more toward data quality, problem framing, feature engineering, automatic algorithm tuning and ensemble methods (combining predictive models), with the algorithms themselves taking more of a backseat role.

Does that make sense?



Andy November 3, 2016 at 10:36 pm #

REPLY ↩

It does make sense.

You mentioned 'data quality'. That's currently my field of work. I've been doing this statistically until now, and very keen to try a different approach. As a practical example how would you use ML to spot an error/outlier using ML instead of stats?

Let's say I have a large dataset containing trees: each tree record contains a specie, height, location, crown size, age, etc... (ah! suspiciously similar to the iris flowers dataset 😊) Is ML a viable method for finding incorrect data and replace with an "estimated" value? The answer I guess is yes. For species I could use almost an identical method to what you presented here; BUT what about continuous values such as tree height?



Jason Brownlee November 4, 2016 at 9:08 am #

REPLY ↩

Hi Andy,

Maybe "outliers" are instances that cannot be easily predicted or assigned ambiguous predicted probabilities.

Instance values can be "fixed" by estimating new values, but whole instance can also be pulled out if data is cheap.



Shailendra Khadayat October 30, 2016 at 2:23 pm #

REPLY ↩

Awesome work Jason. This was very helpful and expect more tutorials in the future.

Thanks.



Jason Brownlee October 31, 2016 at 5:26 am #

REPLY ↩

I'm glad you found it useful Shailendra.



Shuvam Ghosh November 16, 2016 at 12:13 am #

REPLY ↩

Awesome work. Students need to know how the end results will look like. They need to get motivated to learn and one of the effective means of getting motivated is to be able to see and experience the wonderful end results. Honestly, if i were made to study algorithms and understand them i would get bored. But now since i know what amazing results they give, they will serve as driving forces in me to get into details of it and do more research on it. This is where i hate the orthodox college ways of teaching. First get the theory right then apply. No way. I need to see things first to get motivated.



Jason Brownlee November 16, 2016 at 9:29 am #

REPLY ↩

Thanks Shuvam,

I'm glad my results-first approach gels with you. It's great to have you here.



Puneet November 17, 2016 at 12:08 am #

REPLY ↩

Thanks Jason,

while i am trying to complete this.

```
# Spot Check Algorithms
```

```
models = []
```

```
models.append(('LR', LogisticRegression()))
```

```
models.append(('LDA', LinearDiscriminantAnalysis()))
```

```
models.append(('KNN', KNeighborsClassifier()))
```

```
models.append(('CART', DecisionTreeClassifier()))
```

```
models.append(('NB', GaussianNB()))
```

```
models.append(('SVM', SVC()))
```

```
# evaluate each model in turn
```

```
results = []
```

```
names = []
```

```
for name, model in models:
```

```
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
```

```
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

```
    results.append(cv_results)
```

```
    names.append(name)
```

```
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
```

```
    print(msg)
```

showing below error.-

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
```

```
^
```

IndentationError: expected an indented block-



Jason Brownlee November 17, 2016 at 9:54 am #

REPLY ↩

Hi Puneet, looks like a copy-paste error.

Check for any extra new lines or white space around that line that is reporting the error.



Bram March 10, 2018 at 7:51 am #

REPLY ↩

<https://stackoverflow.com/questions/4446366/why-am-i-getting-indentationerror-expected-an-indented-block>

This solved it for me. Copy code to notepad, replace all tabs with 4 spaces.



Jason Brownlee March 11, 2018 at 6:15 am #

REPLY ↩

Nice work.



Puneet November 17, 2016 at 12:30 am #

REPLY ↩

Thanks Json,

I am new to ML. need your help so i can run this.

as i have followed the steps but when trying to build and evalute 5 model using this.

```
# Spot Check Algorithms
```

```
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

facing below mentioned issue.

File "", line 13

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
^
```

IndentationError: expected an indented block

Kindly help.



Martin November 18, 2016 at 5:18 am #

REPLY ↩

Puneet, you need to indent the block (tab or four spaces to the right). That is the way of building a block in Python



george soilis November 17, 2016 at 10:00 pm #

REPLY ↩

just another Python noob here,sending many regards and thanks to Jason :):)



Jason Brownlee November 18, 2016 at 8:22 am #

REPLY ↩

Thanks george, stick with it!



sergio November 22, 2016 at 3:29 pm #

REPLY ↩

Does this tutorial work with other data sets? I'm trying to work on a small assignment and I want to use python



Jason Brownlee November 23, 2016 at 8:50 am #

REPLY ↩

It should provide a great template for new projects sergio

it should provide a great template for new projects Sergio.



Brian February 28, 2018 at 4:10 am #

REPLY ↩

I tried to use another dataset. I am not sure what I imported, but even after changing the names, I still get the petal stuff as output. All of it. I commented out that part of the code and even then it gives me those old outputs.



Albert November 26, 2016 at 1:55 am #

REPLY ↩

Very Awesome step by step for me ! Even I am beginner of python , this gave me many things about Machine learning ~ supervised ML. Appreciate of your sharing !!



Jason Brownlee November 26, 2016 at 10:38 am #

REPLY ↩

I'm glad to hear that Albert.



Umar Yusuf November 27, 2016 at 4:04 am #

REPLY ↩

Thank you for the step by step instructions. This will go along way for newbies like me getting started with machine learning.



Jason Brownlee November 27, 2016 at 10:21 am #

REPLY ↩

You're welcome, I'm glad you found the post useful Umar.



Mike P November 30, 2016 at 6:29 pm #

REPLY ↩

Hi Jason,

Really nice tutorial. I had one question which has had me confused. Once you chose your best model, (in this instance KNN) you then train a new model to be used to make predictions against the validation set. should one not perform K-fold cross-validation on this model to ensure we don't overfit?

if this is correct how would you implement this, from my understanding `cross_val_score` will not allow one to generate a confusion matrix.

I think this is the only thing that I have struggled with in using scikit learn if you could help me it would be much appreciated?



Jason Brownlee December 1, 2016 at 7:26 am #

REPLY ↩

Hi Mike. No.

Cross-validation is just a method to estimate the skill of a model on new data. Once you have the estimate you can get on with things, like confirming you have not fooled yourself (hold out validation dataset) or make predictions on new data.

The skill you report is the cross val skill with the mean and stdev to give some idea of confidence or spread.

Does that make sense?



Mike December 2, 2016 at 1:30 am #

REPLY ↩

Hi Jason,

Thanks for the quick response. So to make sure I understand, one would use cross validation to get a estimate of the skill of a model (mean of cross val scores) or chose the correct hyper parameters for a particular model.

Once you have this information you can just go ahead and train the chosen model with the full training set and test it against the validation set or new data?



Jason Brownlee December 2, 2016 at 8:17 am #

REPLY ↩

Hi Mike. Correct.

Additionally, if the validation result confirms your expectations, you can go ahead and train the model on all data you have including the validation dataset and then start using it in production.

This is a very important topic. I think I'll write a post about it.



Sahana Venkatesh November 30, 2016 at 8:15 pm #

REPLY ↩

This is amazing 🙌 You boosted my morale



Jason Brownlee December 1, 2016 at 7:26 am #

REPLY ↩

I'm so glad to hear that Sahana.



Jhon November 30, 2016 at 8:27 pm #

REPLY ↩

Hi

while doing data visualization and running commands `dataset.plot(.....)` i am having the following error.kindly tell me how to fix it

```
array([[
],
[
]], dtype=object)
```



Jason Brownlee December 1, 2016 at 7:28 am #

REPLY ↩

Looks like no data Jhon. It also looks like it's printing out an object.

Are you running in a notebook or on the command line? The code was intended to be run directly (e.g. command line).



Brendon A. Kay December 1, 2016 at 4:20 am #

REPLY ↩

Hi Jason,

Great tutorial. I am a developer with a computer science degree and a heavy interest in machine learning and mathematics, although I don't quite have the academic background for the latter except for what was required in college. So, this website has really sparked my interest as it has allowed me to learn the field in sort of the "opposite direction".

I did notice when executing your code that there was a deprecation warning for the `sklearn.cross_validation` module. They recommend switching to `sklearn.model_selection`.

When switching the modules I adjusted the following line...

```
kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
```

to...

```
kfold = model_selection.KFold(n_folds=num_folds, random_state=seed)
```

... and it appears to be working okay. Of course, I had switched all other instances of `cross_validation` as well, but it seemed to be that the `KFold()` method dropped the `n` (number of instances) parameter, which caused a runtime error. Also, I dropped the `num_instances` variable.

I could have missed something here, so please let me know if this is not a valid replacement, but thought I'd share!

Once again, great website!



Jason Brownlee December 1, 2016 at 7:33 am #

REPLY ↩

Thanks for the support and the kind words Brendon. I really appreciate it (you made my day!)

Yes, the API has changed/is changing and your updates to the tutorial look good to me, except I think `n_folds` has become `n_splits`.

I will update this example for the new API very soon.



Brendon A. Kay December 1, 2016 at 8:01 am #

REPLY ↩

👍 Now on to more tutorials for me!



Jason Brownlee December 2, 2016 at 8:11 am #

REPLY ↩

You can access more here Brendon:

<http://machinelearningmastery.com/start-here/>



Doug March 9, 2018 at 5:56 am #

Jason, is everything on your website on that page? or is there another site map?

thanks!

P.S. your code ran flawlessly on my Jupyter Notebook fwiw. Although I did get a different result with SVM coming out on top with 99.1667. So I ran the validation set with SVM and came out with 94 93 93 30 fwiw.



Jason Brownlee March 9, 2018 at 6:29 am #

No, not everything, just a small and useful sample.

Yes, machine learning algorithms are stochastic, learn more here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Doug March 9, 2018 at 6:46 am #

Thanks. I actually just read that article. Very helpful.



Sergio December 1, 2016 at 3:41 pm #

REPLY ↩

I'm still having a little trouble understanding step 5.1. I'm trying to apply this tutorial to a new data set but, when I try to evaluate the models from 5.3 I don't get a result.



Jason Brownlee December 2, 2016 at 8:13 am #

REPLY ↩

What is the problem exactly Sergio?

Step 5.1 should create a validation dataset. You can confirm the dataset by printing it out.

Step 5.3 should print the result of each algorithm as it is trained and evaluated.

Perhaps check for a copy-paste error or something?



sergio December 2, 2016 at 9:13 am #

REPLY ↩

Does this tutorial work the exact same way for other data sets? because I'm not using the Hello World dataset



Jason Brownlee December 3, 2016 at 8:23 am #

REPLY ↩

The project template is quite transferable.

You will need to adapt it for your data and for the types of algorithms you want to test.



Jean-Baptiste Hubert December 11, 2016 at 12:17 am #

REPLY ↩

Hi Sir,

Thank you for the information.

I am currently a student, in Engineering school in France.

I am working on date mining project, indeed, I have a many date (40Go) about the price of the stocks of many companies in the CAC40.

My goal is to predict the evolution of the yields and I think that Neural Network could be useful.

My idea is : I take for X the yields from "t=0" to "t=n" and for Y the yields from "t=1 to t=n" and the program should find a relation between the data.

Is that possible ? Is it a good way in order to predict the evolution of the yield ?

Thank you for your time

Hubert

Jean-Baptiste



Jason Brownlee December 11, 2016 at 5:24 am #

REPLY ↩

Hi Jean-Baptiste, I'm not an expert in finance. I don't know if this is reasonable, sorry.

This post might help with phrasing your time series problem for supervised learning:

<http://machinelearningmastery.com/time-series-forecasting-supervised-learning/>



Ernest Bonat December 15, 2016 at 5:33 pm #

REPLY ↩

Hi Jason,

If I include an new item in the models array as:

```
models.append(('LNR - Linear Regression', LinearRegression()))
```

with the library:

```
from sklearn.linear_model import LinearRegression
```

I got an error in the \sklearn\utils\validation.py", line 529, in check_X_y

```
y = y.astype(np.float64)
```

as:

ValueError: could not convert string to float: 'Iris-setosa'

Let me know best to fix that! As you can see from my code, I would like to include the Linear Regression algorithms in my array model too!

Thank you for your help,

Ernest



Jason Brownlee December 16, 2016 at 5:39 am #

REPLY ↩

Hi Ernest, it is a classification problem. We cannot use LinearRegression.

Try adding another classification algorithm to the list.



oumaima December 9, 2017 at 11:29 am #

REPLY ↩

Hi Jason

Hi Jason,

I am new to ML. need your help so i can run this.

```
>>> from matplotlib import pyplot
```

Traceback (most recent call last):

File "", line 1, in

File "c:\python27\lib\site-packages\matplotlib\pyplot.py", line 29, in

import matplotlib.colorbar

File "c:\python27\lib\site-packages\matplotlib\colorbar.py", line 32, in

import matplotlib.artist as martist

File "c:\python27\lib\site-packages\matplotlib\artist.py", line 16, in

from .path import Path

File "c:\python27\lib\site-packages\matplotlib\path.py", line 25, in

from . import _path, rcParams

'ImportError: DLL load failed: %1 n\x92est pas une application Win32 valide.\n'



Jason Brownlee December 10, 2017 at 5:17 am #

REPLY ↩

Sorry, I have not seen that error before. Perhaps this post will help you setup your environment:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Gokul Iyer December 20, 2016 at 2:29 pm #

REPLY ↩

Great tutorial! Quick question, for the when we create the models, we do models.append(name of algorithm, alogrithm function), is models an array? Because it seems like a dictionary since we have a key-value mapping (algorithm name, and algorithm function). Thank you!



Jason Brownlee December 20, 2016 at 2:47 pm #

REPLY ↩

It is a list of tuples where each tuple contains a string name and a model object.



Sasanka ghosh December 21, 2016 at 4:55 am #

REPLY ↩

Hi Jason /any Gurus ,

Good post and will follow it but my question may be little off track.

Asking this question as i am a data modeller /aspiring data architect.

I i feel as Guru/Gurus you can clarify my doubt. The question is at the end .

In current Data management environment

1. Data architecture /Physical implementation and choosing appropriate tools,back end,storage,no sql, SQL, MPP, sharding, columnar ,scale up/out ,distributed processing etc .
2. In addition to DB based procedural languages proficiency at at least one of the following i.e. Java/Python/Scala etc.
3. Then comes this AI,Machine learning ,neural Networks etc .

My question is regarding point 3 .

I believe those are algorithms which needs deep functional knowledge and years of experience to add any value to business .

Those are independent of data models and it's ,physical implementation and part of Business user domain not data architecture domain .

If i take your above example say now 10k users trying to do the similar kind of thing then points 1 and 2 will be Data architects a domain and point 3 will be business analyst domain . may be point 2 can overlap between them to some extent .

Data Architect need not to be hands on/proficient in algorithms i.e. should have just some basic idea as Data architects job is not to invent business logic but implement the business logic physically to satisfy Business users/Analysts .

Am i correct in my assumption as i find the certain things are nearly mutually exclusive and expectations/benchmarks should be set right?

Regards

sasanka ghosh



Jason Brownlee December 21, 2016 at 8:46 am #

REPLY ↩

Hi Sasanka, sorry, I don't really follow.

Are you able to simplify your question?



Sasanka ghosh December 21, 2016 at 9:25 pm #

REPLY ↩

Hi Jason ,

Many thanks that u bothered to reply .

Tried to rephrase and concise but still it is verbose . apologies for that.

Is it expected from a data architect to be algorithm expert as well as data model/database expert?

Algorithms are business centric as well as specific to particular domain of business most of the times.

Giving u an example i.e. SHORTEST PATH (take it as just an example in making my point)

An organization is providing an app to provide that service .

CAVEAT:Someone may say from comp science dept that it is the basic thing u learn but i feel it is still an algorithm not a data structure .

if we take the above scenario in simplistic term the requirement is as follows

- 1.there will be say million registered users
2. one can say at least 10 % are using the app same time
3. any time they can change their direction as per contingency like a military op so dumping the partial weighted graph to their device is not an option i.e. users will be connected to main server/server cluster.
4. the challenge is storing the spatial data in DB in correct data model .
scale out ,fault tolerance .
- 5.implement the shortest path algo and display it using Python/java/Cipher/Oracle spatial/titan etc dynamically.

My question is can a data architect work on this project who does not know the shortest path algorithm but have sufficient knowledge in other areas but the algo with verbose term provided to him/her to implement ?

I m asking this question as now a days people are offering ready made courses etc i.e. machine learning ,NLP,Data scientist etc. and the scenario is confusing

i feel it is misleading as no one can get expert in science overnight and vice versa.

I feel Algorithms are pure science that is a separate discipline .

But to implement it in large scale Scientists/programmers/architects needs to work in tandem with minimal overlapping but continuous discussion.

Last but not the least if i make some sense what is the learning curve should i follow to try to be a data architect in unstructured data in general

regards

sasanka ghosh



Jason Brownlee December 22, 2016 at 6:35 am #

REPLY ↩

Really this depends on the industry and the job. I cannot give you good advice for the general case.

You can get valuable results without being an expert, this applies to most fields.

Algorithms are a tool, use them as such. They can also be a science, but we practitioners don't have the time.

I hope that helps.



Sasanka ghosh December 22, 2016 at 7:00 pm #

Thanks Jsaon.

I appreciate your time and response .

I just wanted to validate from a real techie/guru like u as the confusion or no perfect answer are being exploited by management/HR to their own advantage and practice use and throw policy or make people sycophants/redundant without

following the basic management principle,

The tech guys except "few geniuses" are always toiling and management is opening the cork, enjoying at the same time.

Regards
sasanka ghosh



Raveen Sachintha December 21, 2016 at 8:51 pm #

REPLY ↩

Hello Jason,

Thank you very much for these tutorials, i am new to ML and i find it very encouraging to do and end to end project to get started with rather than reading and reading without seeing and end, This really helped me..

One question, when i tried this i got the highest accuracy for SVM.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

so i decided to try that out too,,

```
svm = SVC()
```

```
svm.fit(X_train, Y_train)
```

```
prediction = svm.predict(X_validation)
```

these were my results using SVM,

0.933333333333

[[7 0 0]

[0 10 2]

[0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.83 0.91 12

Iris-virginica 0.85 1.00 0.92 11

avg / total 0.94 0.93 0.93 30

I am still learning to read these results, but can you tell me why this happened? why did i get high accuracy for SVM instead of KNN?? have i done anything wrong? or is it possible?



Jason Brownlee December 22, 2016 at 6:33 am #

REPLY ↩

The results reported are a mean estimated score with some variance (spread).

It is an estimate on the performance on new data.

When you apply the method on new data, the performance may be in that range. It may be lower if the method has overfit the training data.

Overfitting is a challenge and developing a robust test harness to ensure we don't fool/mislead ourselves during model development is important work.

I hope that helps as a start.



inzar December 25, 2016 at 7:04 am #

REPLY ↩

i want to buy your book.

i try this tutorial and the result is very awesome

i want to learn from you

thanks....



Jason Brownlee December 26, 2016 at 7:41 am #

REPLY ↩

Thanks inzar.

You can see all of my books and bundles here:
<http://machinelearningmastery.com/products>



lou December 25, 2016 at 7:29 am #

REPLY ↩

Why the leading comma in `X = array[:,0:4]`?



Jason Brownlee December 26, 2016 at 7:42 am #

REPLY ↩

This is Python array notation for [rows,columns]

Learn more about slicing arrays in Python here:
<http://structure.usc.edu/numarray/node26.html>



Thinh December 26, 2016 at 5:05 am #

REPLY ↩

In 1.2 , should warn to install scikit-learn



Jason Brownlee December 26, 2016 at 7:49 am #

REPLY ↩

Thanks for the note.

Please see section 1.1 Install SciPy Libraries where it says:

There are 5 key libraries that you will need to install... sklearn



Tijo L. Peter December 28, 2016 at 10:34 pm #

REPLY ↩

Best ML tutorial for Python. Thank you, Jason.



Jason Brownlee December 29, 2016 at 7:17 am #

REPLY ↩

Thanks!



baso December 29, 2016 at 12:38 am #

REPLY ↩

when i tried run, i have error message" TypeError: Empty 'DataFrame': no numeric data to plot" help me



Jason Brownlee December 29, 2016 at 7:18 am #

REPLY ↩

Sorry to hear that.

Perhaps check that you have loaded the data as you expect and that the loaded values are numeric and not strings. Perhaps print the first few rows: `print(df.head(5))`



baso December 29, 2016 at 1:05 pm #

REPLY ↩



baso December 29, 2016 at 1:00 pm #

thanks very much Jason for your time

it worked. these tutorial very help for me. im new in Machine learning, but may you explain to me about your simple project above?
because i did not see X_test and target

regard in advance



Jason Brownlee December 30, 2016 at 5:49 am #

REPLY ↩

Glad to hear it baso!



Andrea January 5, 2017 at 1:42 am #

REPLY ↩

Thank you for sharing this. I bumped into some installation problems.

Eventually, yo get all dependencies installed on MacOS 10.11.6 I had to run this:

brew install python

pip install --user numpy scipy matplotlib ipython jupyter pandas sympy nose scikit-learn

export PATH=\$PATH:~/Library/Python/2.7/bin



Jason Brownlee January 5, 2017 at 9:21 am #

REPLY ↩

Thanks for sharing Andrea.

I'm a macports guy myself, here's my recipe:

```
1 1. Install XCode and XCode Command Line Tools
2 Use the "Mac App Store" application
3 xcode-select --install
4 2. Install Macports
5 https://guide.macports.org/chunked/installing.macports.html
6 3. Install a SciPy Environment
7 sudo port install py27-numpy py27-scipy py27-matplotlib py27-ipython +notebook py27-pandas py27-sympy py27-nose
8 sudo port select --set py-sympy py27-sympy
9 sudo port select --set cython cython27
10 sudo port select --set ipython py27-ipython
11 sudo port select --set ipython2 py27-ipython
12 sudo port select --set python python27
13 sudo port select --set python2 python27
14 sudo port select --set pip pip27
15 4. Install scikit-learn
16 sudo pip install -U scikit-learn
```



Sohib January 6, 2017 at 6:26 pm #

REPLY ↩

Hi Jason,

I am following this page as a beginner and have installed Anaconda as recommended.

As I am on win 10, I installed Anaconda 4.2.0 For Windows Python 2.7 version (x64) and

I am using Anaconda's Spyder (python 2.7) IDE.

I checked all the versions of libraries (as shown in 1.2 Start Python and Check Versions) and got results like below:

Python: 2.7.12 [Anaconda 4.2.0 (64-bit)] (default, Jun 29 2016, 11:07:13) [MSC v.1500 64 bit (AMD64)]

scipy: 0.18.1

numpy: 1.11.1

matplotlib: 1.5.3

pandas: 0.18.1

sklearn: 0.17.1

At the 2.1 Import libraries section, I imported all of them and tried to load data as shown in

2.2 Load Dataset. But when I run it, it doesn't show an output, instead, there is an error:

Traceback (most recent call last):

File "C:\Users\gachon\.spyder\temp.py", line 4, in

from sklearn import model_selection

ImportError: cannot import name model_selection

Below is my code snippet:

```
import pandas
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
print(dataset.shape)
```

When I delete "from sklearn import model_selection" line I get expected results (150, 5).

Am I missing something here?

Thank you for your time and endurance!



Jason Brownlee January 7, 2017 at 8:23 am #

REPLY ↩

Hi Sohib,

You must have scikit-learn version 0.18 or higher installed.

Perhaps Anaconda has documentation on how to update sklearn?



Sohib January 10, 2017 at 12:15 pm #

REPLY ↩

Thank you for reply.

I updated scikit-learn version to 0.18.1 and it helped.

The error disappeared, the result is shown, but one statement

'import sitecustomize' failed; use -v for traceback

is executed above the result.

I tried to find out why, but apparently I might not find the reason.

Is it going to be a problem in my further steps?

How to solve this?

Thank you in advance!



Jason Brownlee January 11, 2017 at 9:25 am #

REPLY ↩

I'm glad to hear it fixed your problem.

Sorry, I don't know what "import sitecustomize" is or why you need it.



Vishakha January 7, 2017 at 10:10 pm #

REPLY ↩

Can i get the same tutorial with java



Abhinav January 8, 2017 at 8:27 pm #

REPLY ↩

Hi Jason,

Nice tutorial.

In univariate plots, you mentioned about gaussian distribution.

According to the univariate plots, sepat-width had gaussian distribution. You said there are 2 variables having gaussain distribution. Please tell the other.

Thanks



Jason Brownlee January 9, 2017 at 7:49 am #

REPLY ↩

The distribution of the others may be multi-modal. Perhaps a double Gaussian.



Thinh January 13, 2017 at 5:07 am #

REPLY ↩

Hi, Jason. Could you please tell me the reason Why you choose KNN in example above ?



Jason Brownlee January 13, 2017 at 9:16 am #

REPLY ↩

Hi Thinh,

No reason other than it is an easy algorithm to run and understand and good algorithm for a first tutorial.



Scott P January 13, 2017 at 10:25 pm #

REPLY ↩

Hi Jason,

I'm trying to use this code with the KDD Cup '99 dataset, and I am having trouble with LabelEncoding my dataset in to numerical values.

#Modules

import pandas

import numpy

from pandas.tools.plotting import scatter_matrix

import matplotlib.pyplot as plt

from sklearn import preprocessing

from sklearn import cross_validation

from sklearn.metrics import classification_report

from sklearn.metrics import confusion_matrix

from sklearn.metrics import accuracy_score

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn.naive_bayes import GaussianNB

from sklearn.svm import SVC

from sklearn.preprocessing import LabelEncoder

#new

from collections import defaultdict

#

#Load KDD dataset

data_set = "NSL-KDD/KDDTrain+.txt"

names =

```
['duration','protocol_type','service','flag','src_bytes','dst_bytes','land','wrong_fragment','urgent','hot','num_failed_logins','logged_in','num_compromised','su_attempted','num_root','num_file_creations','num_shells','num_access_files','num_outbound_cmds','is_host_login','is_guest_login','count','srv_count','error_rate','srv_error_rate','rerror_rate','srv_error_rate','same_srv_rate','diff_srv_rate','srv_diff_host_rate','dst_host_count','dst_host_srv_count','dst_host_same_srv_rate','dst_host_diff_srv_rate','dst_host_same_src_port_rate','dst_host_srv_diff_host_rate','dst_host_error_rate','dst_host_srv_error_rate','dst_host_error_rate','dst host srv error rate','class']
```

```

#Diabetes Dataset
#data_set = "Datasets/pima-indians-diabetes.data"
#names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
#data_set = "Datasets/iris.data"
#names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
dataset = pandas.read_csv(data_set, names=names)

array = dataset.values
X = array[:,0:40]
Y = array[:,40]

label_encoder = LabelEncoder()
label_encoder = label_encoder.fit(Y)
label_encoded_y = label_encoder.transform(Y)

validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, label_encoded_y, test_size=validation_size,
random_state=seed)

# Test options and evaluation metric
num_folds = 7
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'

# Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean()*100, cv_results.std()*100)#multiplying by 100 to show percentage
    print(msg)

# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(Y)
plt.show()

```

Am I doing something wrong with the LabelEncoding process?



MegO_Bonus June 4, 2017 at 7:15 pm #

REPLY ↩

Hi. Change all symbols like " to " and ' to '. LabaleEncoder will be work correct but not all network. I try to create a neural network for NSL-KDD too. Have you any good examples?



Jason Brownlee June 5, 2017 at 7:40 am #

REPLY ↩

What is "NSL-KDD"?



bugtime December 10, 2017 at 8:22 pm #

REPLY ↩

Hello Jason,

Please see https://github.com/defcom17/NSL_KDD



Jason Brownlee December 11, 2017 at 5:25 am #

I'm not familiar with this, sorry.



Dan January 14, 2017 at 4:56 am #

REPLY ↩

Hi, I'm running a bit of a different setup than yours.

The modules and version of python I'm using are more recent releases:

Python: 3.5.2 |Anaconda 4.2.0 (32-bit)| (default, Jul 5 2016, 11:45:57) [MSC v.1900 32 bit (Intel)]

scipy: 0.18.1

numpy: 1.11.3

matplotlib: 1.5.3

pandas: 0.19.2

sklearn: 0.18.1

And I've gotten SVM as the best algorithm in terms of accuracy at 0.991667 (0.025000).

Would you happen to know why this is, considering more recent versions?

I also happened to get a rather different boxplot but I'll leave it at what I've said thus far.



Jason Brownlee January 15, 2017 at 5:26 am #

REPLY ↩

Hi Dan,

You may get differing results for a variety of reasons. Small changes in the code will affect the result. This is why we often report mean and stdev algorithm performance rather than one number, to given a range of expected performance.

This post on randomness in ml algorithms might also help:

<http://machinelearningmastery.com/randomness-in-machine-learning/>



Duncan Carr January 17, 2017 at 1:44 am #

REPLY ↩

Hi Jason

I can't tell you how grateful I am ... I have been trawling through lots of ML stuff to try to get started with a "toy" example. Finally I have found the tutorial I was looking for. Anaconda had old sklearn: 0.17.1 for Windows – which caused an error "ImportError: cannot import name 'model_selection'". That was fixed by running "pip install -U scikit-learn" from the Anaconda command-line prompt. Now upgraded to 0.18. Now everything in your imports was fine.

All other tutorials were either too simple or too complicated. Usually the latter!

Thank you again 😊



Jason Brownlee January 17, 2017 at 7:39 am #

REPLY ↩

Glad to hear it Duncan.

Thanks for the tip for Anaconda uses.

I'm here to help if you have questions!



Malathi January 17, 2017 at 3:13 am <#>

REPLY

Hi Jason,

Wonderful service. All of your tutorials are very helpful to me. Easy to understand.

Expecting more tutorials on deep neural networks.

Malathi



Jason Brownlee January 17, 2017 at 7:40 am <#>

REPLY

You're very welcome Malathi, glad to hear it.



Duncan Carr January 17, 2017 at 7:32 pm <#>

REPLY

Hi Jason

I managed to get it all working – I am chuffed to bits.

I get exactly the same numbers in the classification report as you do ... however, when I changed both seeds to 8 (from 7), then ALL of the numbers end up being 1. Is this good, or bad? I am a bit confused.

Thanks again.



Jason Brownlee January 18, 2017 at 10:14 am <#>

REPLY

Well done Duncan!

What do you mean all the numbers end up being one?



Duncan Carr January 18, 2017 at 8:02 pm <#>

REPLY

Hi Jason

I've output the "accuracy_score", "confusion_matrix" & "classification_report" for seeds 7, 9 & 10. Why am I getting a perfect score with seed=9? Many thanks.

(seed=7)

0.9

```
[[10 0 0]
```

```
[ 0 8 1]
```

```
[ 0 2 9]]
```

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 10

Iris-versicolor 0.80 0.89 0.84 9

Iris-virginica 0.90 0.82 0.86 11

avg / total 0.90 0.90 0.90 30

(seed=9)

1.0

```
[[13 0 0]
```

```
[ 0 9 0]
```

```
[ 0 0 8]]
```

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 13

Iris-versicolor 1.00 1.00 1.00 9

Iris-virginica 1.00 1.00 1.00 8

avg / total 1.00 1.00 1.00 30

(seed=10)

0.9666666666666667

[[10 0 0]

[0 12 1]

[0 0 7]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 10

Iris-versicolor 1.00 0.92 0.96 13

Iris-virginica 0.88 1.00 0.93 7

avg / total 0.97 0.97 0.97 30



Jason Brownlee January 19, 2017 at 7:31 am #

REPLY ↩

Random chance. This is why it is a good idea to use cross-validation with many repeats and report mean and standard deviation scores.

More on randomness in machine learning here:

<http://machinelearningmastery.com/randomness-in-machine-learning/>



shivani January 20, 2017 at 8:40 pm #

REPLY ↩

from sklearn import model_selection
showing Import Error: can not import model_selection



Jason Brownlee January 21, 2017 at 10:25 am #

REPLY ↩

You need to update your version of sklearn to 0.18 or higher.



Jim January 22, 2017 at 5:06 pm #

REPLY ↩

Jason

Excellent Tutorial. New to Python and set a New Years Resolution to try to understand ML. This tutorial was a great start.

I struck the issue of the sklearn version. I am using Ubuntu 16.04 LTS which comes with python-sklearn version 0.17. To update to latest I used the site:

http://neuro.debian.net/install_pkg.html?p=python-sklearn

Which gives the commands to add the neuro repository and pull down the 0.18 version.

Also I would like to note there is an error in section 3.1 Dimensions of the Dataset. Your text states 120 Instances when in fact 150 are returned, which you have in the Printout box.

Keep up the good work.

Jim



Jason Brownlee January 23, 2017 at 8:37 am #

REPLY ↩

I'm glad to hear you worked around the version issue Jim, nice work!

Thanks for the note on the typo, fixed!



Raphael January 23, 2017 at 4:15 pm #

REPLY ↩

hi Jason.nice work here. I'm new to your blog. What does the y-axis in the box plots represent?



Jason Brownlee January 24, 2017 at 11:01 am #

REPLY ↩

Hi Raphael,

The y-axis in the box-and-whisker plots are the scale or distribution of each variable.



Kayode January 23, 2017 at 8:42 pm #

REPLY ↩

Thank you for this wonderful tutorial.



Jason Brownlee January 24, 2017 at 11:03 am #

REPLY ↩

You're welcome Kayode.



Raphael January 26, 2017 at 2:28 am #

REPLY ↩

hi Jason,

In this line

```
dataset.groupby('class').size()
```

what other variable other than size could I use? I changed size with count and got something similar but not quite. I got key errors for the other stuffs I tried. Is size just a standard command?



Jason Brownlee January 26, 2017 at 4:48 am #

REPLY ↩

Great question Raphael.

You can learn more about Pandas groupby() here:

<http://pandas.pydata.org/pandas-docs/stable/groupby.html>



Scott January 26, 2017 at 10:35 pm #

REPLY ↩

Jason,

I'm trying to use a different data set (KDD CUP 99') with the above code, but when I try and run the code after modifying "names" and the array to account for the new features and it will not run as it is giving me an error of: "cannot convert string to a float".

In my data set, there are 3 columns that are text and the rest are integers and floats, I have tried LabelEncoding but it gives me the same error, do you know how I can resolve this?



Jason Brownlee January 27, 2017 at 12:08 pm #

REPLY ↩

Hi Scott,

If the values are indeed strings, perhaps you can use a method that supports strings instead of numbers, perhaps like a decision tree.

If there are only a few string values for the column, a label encoding as integers may be useful.

Alternatively, perhaps you could try removing those string features from the dataset.

I hope that helps, let me know how you go.



Weston Green January 27, 2017 at 12:11 pm #

REPLY ↩



weston GROSS January 31, 2017 at 10:41 am #

I would like a chart to see the grand scope of everything for data science that python can do.

You list 6 basic steps. For example in the visualizing step, I would like to know what all the charts are, what they are used for, and what python library it comes from.

I am extremely new to all this, and understand that some steps have to happen for example

1. Get Data
2. Validate Data
3. Missing Data
4. Machine Learning
5. Display Findinds

So for missing data, there are techniques to restore the data, what are they and what libraries are used?



Jason Brownlee February 1, 2017 at 10:36 am #

REPLY ↩

You can handle missing data in a few ways such as:

1. Remove rows with missing data.
2. Impute missing data (e.g. use the Imputer class in sklearn)
3. Use methods that support missing data (e.g. decision trees)

I hope that helps.



Mohammed February 1, 2017 at 1:11 am #

REPLY ↩

Hi Jason,

I am a Non Tech Data Analyst and use SPSS extensively on Academic / Business Data over the last 6 years.

I understand the above example very easily.

I want to work on Search – Language Translation and develop apps.

Whats the best way forward ...

Do you also provide Skype Training / Project Mentoring..

Thanks in advance.



Jason Brownlee February 1, 2017 at 10:51 am #

REPLY ↩

Thanks Mohammed.

Sorry, I don't have good advice for language translation applications.



Mohammed February 1, 2017 at 1:14 am #

REPLY ↩

I dont have any Development / Coding Background.

However, following your guidelines I downloaded SciPy and tested the code.

Everything worked perfectly fine.

Looking forward to go all in...



Jason Brownlee February 1, 2017 at 10:51 am #

REPLY ↩

I'm glad to hear that Mohammed



Jason Brownlee February 1, 2017 at 10:51 am #

REPLY ↩



Purvi February 1, 2017 at 7:31 am #

Hi Jason,

I am new to Machine learning and am trying out the tutorial. I have following environment :

```
>>> import sys
>>> print('Python: {}'.format(sys.version))
Python: 2.7.10 (default, Jul 13 2015, 12:05:58)
[GCC 4.2.1 Compatible Apple LLVM 6.1.0 (clang-602.0.53)]
>>> import scipy
>>> print('scipy: {}'.format(scipy.__version__))
scipy: 0.18.1
>>> import numpy
>>> print('numpy: {}'.format(numpy.__version__))
numpy: 1.12.0
>>> import matplotlib
>>> print('matplotlib: {}'.format(matplotlib.__version__))
matplotlib: 2.0.0
>>> import pandas
>>> print('pandas: {}'.format(pandas.__version__))
pandas: 0.19.2
>>> import sklearn
>>> print('sklearn: {}'.format(sklearn.__version__))
sklearn: 0.18.1
```

When I try to load the iris dataset, it loads up fine and prints dataset.shape but then my python interpreter hangs. I tried it out 3-4 times and everytime it hangs after I run couple of commands on dataset.

```
>>> url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
>>> names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
>>> dataset = pandas.read_csv(url, names=names)
>>> print(dataset.shape)
(150, 5)
>>> print(dataset.head(20))
sepal-length sepal-width petal-length petal-width class
0 5.1 3.5 1.4 0.2 Iris-setosa
1 4.9 3.0 1.4 0.2 Iris-setosa
2 4.7 3.2 1.3 0.2 Iris-setosa
3 4.6 3.1 1.5 0.2 Iris-setosa
4 5.0 3.6 1.4 0.2 Iris-setosa
5 5.4 3.9 1.7 0.4 Iris-setosa
6 4.6 3.4 1.4 0.3 Iris-setosa
7 5.0 3.4 1.5 0.2 Iris-setosa
8 4.4 2.9 1.4 0.2 Iris-setosa
9 4.9 3.1 1.5 0.1 Iris-setosa
10 5.4 3.7 1.5 0.2 Iris-setosa
11 4.8 3.4 1.6 0.2 Iris-setosa
12 4.8 3.0 1.4 0.1 Iris-setosa
13 4.3 3.0 1.1 0.1 Iris-setosa
14 5.8 4.0 1.2 0.2 Iris-setosa
15 5.7 4.4 1.5 0.4 Iris-setosa
16 5.4 3.9 1.3 0.4 Iris-setosa
17 5.1 3.5 1.4 0.3 Iris-setosa
18 5.7 3.8 1.7 0.3 Iris-setosa
19 5.1 3.8 1.5 0.3 Iris-setosa
>>> print(dataset)
```

It does not let me type anything further.

I would appreciate your help.


Thanks,

Purvi



Jason Brownlee February 1, 2017 at 10:55 am #

Hi Purvi, sorry to hear that.

REPLY 

Perhaps you're able to comment out the first parts of the tutorial and see if you can progress?



sam February 5, 2017 at 9:24 am #

REPLY ↩

Hi Jason

i am planning to use python to predict customer attrition.I have current list of attrited customers with their attributes.I would like to use them as test data and use them to predict any new customers.Can you please help to approach the problem in python ?

my test data :

customer1 attribute1 attribute2 attribute3 ... attrited

my new data

customer N, attribute 1,..... ?

Thanks for your help in advance.



Jason Brownlee February 6, 2017 at 9:42 am #

REPLY ↩

Hi Sam, as a start, this process will help you clearly define and work through your predictive modeling problem:

<http://machinelearningmastery.com/start-here/#process>

I'm happy to answer questions as you work through the process.



Kiran Prajapati February 7, 2017 at 6:31 pm #

REPLY ↩

Hello Sir, I want to check my data is how many % accurate, In my data , I have 4 columns ,

Taluka , Total_yield, Rain(mm) , types_of soil

Nasik 12555 63.0 dark black

Igatpuri 1560 75.0 shallow

So on,

first, I have to check data is accurate or not, and next step is what is the predicted yield , using regression model.

Here is my model Total_yield = Rain + types_of soil

I use 0 and 1 binary variable for types_of soil.

can you please help me, how to calculate data is accurate ? How many % ?

and how to find predicted yield ?



Jason Brownlee February 8, 2017 at 9:33 am #

REPLY ↩

I'm not sure I understand Kiran.

This process will help you describe and work through your predictive modeling project:

<http://machinelearningmastery.com/start-here/#process>



Saby February 15, 2017 at 9:11 am #

REPLY ↩

Load dataset

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
```

```
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
```

```
dataset = pandas.read_csv(url, names=names)
```

The dataset should load without incident.

If you do have network problems, you can download the iris.data file into your working directory and load it using the same method, changing url to the local file name.

I am a very beginner python learner(trying to learn ML as well), I tried to load data from my local file but could not be successful. Will you help me out how exactly code should be written to open the data from local file.

help me confirm exactly what needs to happen to open the data from local files.



Jason Brownlee February 15, 2017 at 11:39 am #

REPLY ↩

Sure.

Download the file as iris.data into your current working directory (where your python file is located and where you are running the code from).

Then load it as:

```
1 dataset = pandas.read_csv('iris.data', names=names)
```



ant February 15, 2017 at 9:54 pm #

REPLY ↩

Hi, Jason, first of all thank so much for this amazing lesson.

Just for curiosity I have computed all the values obtained with dataset.describe() with excel and for the 25% value of petal-length I get 1.57500 instead of 1.60000. I have googled for formatting describe() output unsuccessfully. Is there an explanation? Tnx



Jason Brownlee February 16, 2017 at 11:07 am #

REPLY ↩

Not sure, perhaps you could look into the Pandas source code?



ant February 17, 2017 at 12:23 am #

REPLY ↩

OK, I will do.



jacques February 16, 2017 at 4:42 pm #

REPLY ↩

Hi Jason

I don't quite follow the KFOLD section ?

We started of with 150 data-entries(rows)

We then use a 80/20 split for validation/training that leaves us with 120

The split 10 boggles me ??

Does it take 10 items from each class and train with 9 ? what does the other 1 left do then ?



Jason Brownlee February 17, 2017 at 9:52 am #

REPLY ↩

Hi jacques,

The 120 records are split into 10 folds. The model is trained on the first 9 folds and evaluated on the records in the 10th. This is repeated so that each fold is given a chance to be the hold out set. 10 models are trained, 10 scores collected and we report the mean of those scores as an estimate of the performance of the model on unseen data.

Does thar help?



Alhassan February 17, 2017 at 4:02 pm #

REPLY ↩

I am trying to integrate machine learning into a PHP website I have created. Is there any way I can do that using the guidelines you provided above?



Jason Brownlee February 18, 2017 at 8:34 am #

REPLY ↩



I have not done this Alhassan.

Generally, I would advise developing a separate service that could be called using REST calls or similar.

If you are working on a prototype, you may be able to call out to a program or script from cgi-bin, but this would require careful engineering to be secure in a production environment.



Simão Gonçalves February 20, 2017 at 1:27 am #

REPLY ↩

Hi Jason! This tutorial was a great help, i'm truly grateful for this so thank you.

I have one question about the tutorial though, in the Scattplot Matrix i can't understand how can we make the dots in the graphs whose variables have no relationship between them (like sepal-length with petal_width).

Could you or someone explain that please? how do you make a dot that represents the relationship between a certain sepal_length with a certain petal-width



Jason Brownlee February 20, 2017 at 9:30 am #

REPLY ↩

Hi Simão,

The x-axis is taken for the values of the first variable (e.g. sepal_length) and the y-axis is taken for the second variable (e.g. petal_width).

Does that help?



Yopo February 21, 2017 at 4:35 am #

REPLY ↩

you match each iris instance's length and width with each other. for example, iris instance number one is represented by a dot, and the dot's values are the iris length and width! so actually, when you take all these values and put them on a graph you are basically checking to see if there is a relation. as you can see some in some of these plots the dots are scattered all around, but when we look at the petal width – petal length graph it seems to be linear! this means that those two properties are clearly related. hope this helped!



Sébastien February 20, 2017 at 9:34 pm #

REPLY ↩

Hi Jason,

from France and just to say you "Thank you for this very clear tutorial!"

Sébastien



Jason Brownlee February 21, 2017 at 9:34 am #

REPLY ↩

I'm glad you found it useful Sébastien.



Raj February 27, 2017 at 2:53 am #

REPLY ↩

Hi Jason,

I am new to ML & Python. Your post is encouraging and straight to the point of execution. Anyhow, I am facing below error when

```
>>> validation_size = 0.20
>>> X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state = seed)
Traceback (most recent call last):
  File "", line 1, in
NameError: name 'validation_size' is not defined
```

What could be the miss out? I didn't get any errors in previous steps.

My Environment details:

OS: Windows 10

OS: windows 10
Python : 3.5.2
scipy : 0.18.1
numpy : 1.11.1
sklearn : 0.18.1
matplotlib : 0.18.1



Jason Brownlee February 27, 2017 at 5:54 am #

REPLY ↩

Hi Raj,

Double check you have the code from section "5.1 Create a Validation Dataset" where validation_size is defined.

I hope that helps.



Roy March 2, 2017 at 7:38 am #

REPLY ↩

Hey Jason,

Can you please explain what precision, recall, f1-score, support actually refer to?

Also what the numbers in a confusion matrix refers to?

[7 0 0]

[0 11 1]

[0 2 9]]

Thanks.



Jason Brownlee March 2, 2017 at 8:24 am #

REPLY ↩

Hi Roy,

You can learn all about the confusion matrix in this post:

<http://machinelearningmastery.com/confusion-matrix-machine-learning/>

You can learn all about precision and recall in this article:

https://en.wikipedia.org/wiki/Precision_and_recall



Ahmed December 25, 2017 at 2:21 am #

REPLY ↩

Hi Jason,

Thank you very much for your tutorial.

I am a little bit confused about the confusion matrix, because you are using a 3×3 matrix while it should be a 2×2 matrix.



Jason Brownlee December 25, 2017 at 5:25 am #

REPLY ↩

Learn more about the confusion matrix here:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



Ahmed December 25, 2017 at 6:16 am #

REPLY ↩

Hi Jason,

Now I understand the meaning of your confusion matrix, so I don't need any explanation.

Thank you and best regards.



Jason Brownlee December 26, 2017 at 5:11 am #

REPLY ↩

You're welcome.



santosh March 3, 2017 at 7:29 am #

REPLY ↩

what code should i use to load data from my working directory??



Jason Brownlee March 3, 2017 at 7:47 am #

REPLY ↩

This post will help you out Santosh:

<http://machinelearningmastery.com/load-machine-learning-data-python/>



David March 7, 2017 at 8:27 am #

REPLY ↩

Hi Jason,

I have a ValueError and i don't know how can i solve this problem

My problem like that,

ValueError: could not convert string to float: '2013-06-27 11:30:00.0000000'

Can u give some information about the fixing this problem?

Thank you



Jason Brownlee March 7, 2017 at 9:39 am #

REPLY ↩

It looks like you are trying to load a date-time. You might need to write a custom function to parse the date-time when loading or try removing this column from your dataset.



Saugata De March 8, 2017 at 6:11 am #

REPLY ↩

```
>>> for name, model in models:
```

```
... kfold=model_selection.Kfold(n_splits=10, random_state=seed)
... cv_results =model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
... results.append(cv_results)
... names.append(name)
... msg="%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
... print(msg)
...
```

After typing this piece of code, it is giving me this error. can you plz help me out Jason. Since I am new to ML, dont have so much idea about the error.

Traceback (most recent call last):

File "", line 2, in

AttributeError: module 'sklearn.model_selection' has no attribute 'Kfold'



Asad Ali July 23, 2017 at 12:59 pm #

REPLY ↩

the KFold function is case-sensitive. It is " model_selection.KFold(...)" not " model_selection.Kfold(...)"
update this line:

```
kfold=model_selection.KFold(n_splits=10, random_state=seed)
```



ibtssam February 12, 2018 at 9:17 pm #

REPLY ↩

THANK U



Ojas March 10, 2017 at 10:58 am #

REPLY ↩

Hello Jason ,

Thanks for writing such a nice and explanatory article for beginners like me but i have one concern , i tried finding it out on other websites as well but could not come up with any solution.

Whatever i am writing inside the code editor (Jupyter Qtconsole in my case) , can this not be save as a .py file and shared with my other members over github maybe?. I found some hacks though but i have a thinking that there must be some proper way of sharing the codes written in the editor. , like without the outputs or plots in between.



Jason Brownlee March 11, 2017 at 7:55 am #

REPLY ↩

You can write Python code in a text editor and save it as a myfile.py file. You can then run it on the command line as follows:

```
1 python myfile.py
```

Consider picking up a book on Python.



manoj maracheea March 11, 2017 at 9:37 pm #

REPLY ↩

Hello Jason,

Nice tutorials I done this today.

I didn't really understand everything, { I will follow your advice, will do it again, write all the question down, and use the help function.}

The tutorials just works, I take around 2 hours to do it typing every single line.
install all the dependencies, run on each blocks types, to check.

Thanks, I be visiting your blogs, time to time.

Regards,



Jason Brownlee March 12, 2017 at 8:23 am #

REPLY ↩

Well done, and thanks for your support.

Post any questions you have as comments or email me using the "contact" page.



manoj maracheea March 11, 2017 at 9:38 pm #

REPLY ↩

Just I am a beginner too, I am using Visual studio code.

Look good.



Vignesh R March 13, 2017 at 9:59 pm #

REPLY ↩

What exactly is confusion matrix?



Jason Brownlee March 14, 2017 at 8:18 am #

REPLY ↩

Great question, see this post:

<http://machinelearningmastery.com/confusion-matrix-machine-learning/>



Dan R. March 14, 2017 at 7:09 am #

REPLY ↩

Can I ask what is the reason of this problem? Thank for answer 🙏 :
(In my code is just the section, where I Import all the needed libraries..)
I have all libraries up to date, but it still gives me this error->

File "C:\Users\64dri\Anaconda3\lib\site-packages\sklearn\model_selection_search.py", line 32, in
from ..utils.fixes import rankdata

ImportError: cannot import name 'rankdata'

(scipy: 0.18.1
numpy: 1.11.1
matplotlib: 1.5.3
pandas: 0.18.1
sklearn: 0.17.1)



Jason Brownlee March 14, 2017 at 8:31 am #

REPLY ↩

Sorry, I have not seen this issue Dan, consider searching or posting to StackOverflow.



Cameron March 15, 2017 at 5:28 am #

REPLY ↩

Jason,

You're a rockstar, thank you so much for this tutorial and for your books! It's been hugely helpful in getting me started on machine learning. I was curious, is it possible to add a non-number property column, or will the algorithms only accept numbers?

For example, if there were a "COLOR" column in the iris dataset, and all Iris-Setosa were blue. how could I get this program to accept and process that COLOR column? I've tried a few things and they all seem to fail.



Jason Brownlee March 15, 2017 at 8:16 am #

REPLY ↩

Great question Cameron!

sklearn requires all input data to be numbers.

You can encode labels like colors as integers and model that.

Further, you can convert the integers to a binary encoding/one-hot encoding which may be more suitable if there is no ordinal relationship between the labels.



Cameron March 15, 2017 at 2:19 pm #

REPLY ↩

Jason, thanks so much for replying! That makes a lot of sense. When you say binary/one-hot encoding I assume you mean (continuing to use the colors example) adding a column for each color (R,O,Y,G,B,V) and for each flower putting a 1 in the column of it's color and a 0 for all of the other colors?

That's feasible for 6 colors (adding six columns) but how would I manage if I wanted to choose between 100 colors or 1000 colors? Are there other libraries that could help deal with that?



Jason Brownlee March 16, 2017 at 7:58 am #

REPLY ↩

Yes you are correct.

Yes, sklearn offers LabelEncoder and OneHotEncoder classes.

Here is a tutorial to get you started:

<http://machinelearningmastery.com/data-preparation-gradient-boosting-xgboost-python/>



Cameron March 19, 2017 at 3:50 am #

Awsome! thanks so much Jason!

AWESOME! THANKS SO MUCH JASON!



Jason Brownlee March 19, 2017 at 9:11 am #

You're welcome, let me know how you go.



James March 19, 2017 at 6:54 am #

REPLY ↩

```
for name, model in models:
... kfold = cross_validation.KFold(n=num_instances,n_folds=num_folds,random_state=seed)
... cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
File "", line 3
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
^
SyntaxError: invalid syntax
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> cv_results = model_selection.cross_val_score(models, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'kfold' is not defined
>>> cv_results = model_selection.cross_val_score(models, X_train, Y_train, cv =
kfold, scoring = scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'kfold' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
```

I am new to python and getting these errors after running 5.3 models. Please help me.



Jason Brownlee March 19, 2017 at 9:12 am #

REPLY ↩

It looks like you might not have copied all of the code required for the example.



Mier March 20, 2017 at 10:26 am #

REPLY ↩

Hi, I went through your tutorial. It is super great!

I wonder whether you can recommend a data set that is similar to Iris classification for me to practice?



Jason Brownlee March 21, 2017 at 8:36 am #

REPLY ↩

Thanks Mier,

I recommend some datasets here:

<http://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/>



Medine H. March 23, 2017 at 2:56 am #

REPLY ↩

Hi Jason,

That's an amazing tutorial, quite clear and useful.

Thanks a bunch!



Jason Brownlee March 23, 2017 at 8:50 am #

REPLY ↩

Thanks Medine.



Sean March 23, 2017 at 9:54 am #

REPLY ↩

Hi Jason,

Can you let me know how can I start with Fraud Detection algorithms for a retail website ?

Thanks,
Sean



Jason Brownlee March 24, 2017 at 7:51 am #

REPLY ↩

Hi Sean, this process will help you work through your problem:

<http://machinelearningmastery.com/start-here/#process>



Raja March 24, 2017 at 11:08 am #

REPLY ↩

You are doing great with your work.

I need your suggestion, i am working on my thesis here i need to work on machine learning.

Training : positive ,negative, others

Test : unknown data

Want to train machine with training and test with unknown data using SVM,Naive,KNN

How can i make the format of training and test data ?

And how to use those algorithms in it

Using which i can get the TP,TN,FP,FN

Thanking you..



Jason Brownlee March 25, 2017 at 7:31 am #

REPLY ↩

This article might help:

https://en.wikipedia.org/wiki/Precision_and_recall



Sey March 26, 2017 at 12:38 am #

REPLY ↩

I m new in Machine learning and this was a really helpful tutorial. I have maybe a stupid question I wanted to plot the predictions and the validation value and make a visual comparison and it doesn't seem like I really understood how I can plot it.

Can you please send me the piece of code with some explanations to do it ?

thank you very much



Jason Brownlee March 26, 2017 at 6:13 am #

REPLY ↩

You can use matplotlib, for example:

```
1 yhat = model.predict(X)
2 from matplotlib import pyplot
3 pyplot.plot(y, yhat)
4 pyplot.show()
```



Kamol Roy March 26, 2017 at 7:25 am #

REPLY ↩

Thanks a lot. It was very helpful.



Jason Brownlee March 27, 2017 at 7:51 am #

REPLY ↩

You're welcome Kamol, I'm glad to hear it.



Rajneesh March 29, 2017 at 11:31 pm #

REPLY ↩

Hi

Sorry for a dumb question.

Can you briefly describe, what the end result means (i.e.. what the program has predicted)



Jason Brownlee March 30, 2017 at 8:53 am #

REPLY ↩

Given an input description of flower measurements, what species of flower is it?

We are predicting the iris flower species as one of 3 known species.



Anusha Vidapanakal March 30, 2017 at 3:58 am #

REPLY ↩

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Why am I getting the highest accuracy for SVM?

I'm a beginner, there was a similar query above but I couldn't quite understand your reply.

Could you please help me out? Have I done any mistake?



Jason Brownlee March 30, 2017 at 8:56 am #

REPLY ↩

Why is a very hard question to answer.

Our role is to find what works, ensure the results are robust, then figure out how we can use the model operationally.



Anusha Vidapanakal March 30, 2017 at 11:33 pm #

REPLY ↩

Okay. Thanks a lot for the prompt response!

The tutorial was very helpful.



Jason Brownlee March 31, 2017 at 5:54 am #

REPLY ↩

Glad to hear it Anusha.



Vinay March 31, 2017 at 11:10 pm #

REPLY ↩

Great tutorial Jason!

My question is, if I want some new data from a user, how do I do that? If in future I develop my own machine learning algorithm, how do I use it to get some new data?

What steps are taken to develop it?

And thanks for this tutorial.



Jason Brownlee April 1, 2017 at 5:56 am #

REPLY ↩

Not sure I understand. Collect new data from your domain and store it in a CSV or write code to collect it.



walid barakeh April 2, 2017 at 6:31 pm #

REPLY ↩

Hi Jason,

I have a question regards the step after trained the data and know the better algorithm for our case, how we could know the rules formula that the algorithm produced for future uses ?

and thanks for the tutorial, its really helpful



Jason Brownlee April 4, 2017 at 9:06 am #

REPLY ↩

You can extract the weights if you like. Not sure I understand why you want the formula for the network. It would be complex and generally unreadable.

You can finalize the mode, save the weights and topology for later use if you like.



walid barakeh April 5, 2017 at 7:40 pm #

REPLY ↩

the best algorithm results for my use case was the "Classification and Regression Trees (CART)", so how could I know the rules that the algorithm created on my usecase.

how I could extract the weights and use them for evaluate new data .

Thanks for your prompt response



Jason Brownlee April 9, 2017 at 2:34 pm #

REPLY ↩

See this post on how to finalize your model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>



Divya April 4, 2017 at 4:58 pm #

REPLY ↩

Thank you so much...this document really helped me a lot.....i was searching for such a document since a long time...this document gave the actual view of how machine learning is implemented through python....Books and courses are really difficult to understand completely and begin with development of project on such a vast concept... books n videos gave me lots of snippets, but i was not understanding how they all fit together.



Jason Brownlee April 9, 2017 at 2:30 pm #

REPLY ↩

I'm glad to hear that.



Divya April 4, 2017 at 5:00 pm #

REPLY ↩

can i get such more tutorials for more detailed understanding?.....It will be really helpfull.



Jason Brownlee April 9, 2017 at 2:30 pm #

REPLY ↩

Sure, see here:

<http://machinelearningmastery.com/start-here/#python>



Gav April 11, 2017 at 5:17 pm #

REPLY ↩

Can't load the iris dataset either through the url or copied to working folder without the NameError: name 'pandas' is not defined



Jason Brownlee April 12, 2017 at 7:51 am #

REPLY ↩

You need to install the Pandas library.

See this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Gavin April 12, 2017 at 9:53 pm #

REPLY ↩

I've already installed Anaconda with Python 3.6 and the panda libraries are listed when I run versions.py. Everything has been fine up till trying to load the iris library. Do I need to use a different terminal within Anaconda?



Jason Brownlee April 13, 2017 at 10:01 am #

REPLY ↩

You may need to close and re-open the terminal window, or maybe restart your system after installation.



Sunil June 4, 2017 at 2:31 am #

REPLY ↩

add a line

import pandas
at the top



Jason Brownlee June 4, 2017 at 7:54 am #

Thanks Sunil!



Ursula April 13, 2017 at 7:33 pm #

REPLY ↩

Hi Jason,

Your tutorial is fantastic!

I'm trying to follow it but gets stuck on 5.3 Build Models

When I copy your code for this section I get a few Errors

IndentationError: expected an indented block

NameError: name 'model' is not defined

NameError: name 'cv_results' is not defined

NameError: name 'name' is not defined

Could you please help me find what I'm doing wrong?

Thanks!

see the code and my "results" below:

```
>>> # Spot Check Algorithms
```

```

... models = []
>>> models.append(('LR', LogisticRegression()))
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)

```



Jason Brownlee April 14, 2017 at 8:43 am #

REPLY ↩

Make sure you have the same tab indenting as in the example. Maybe re-add the tabs yourself after you copy-paste the code.



Nathan Wilson March 26, 2018 at 11:16 am #

REPLY ↩

I'm having this same problem. How would I add the Indentations after I paste the code? Whenever I paste the code, it automatically executes the code.



Jason Brownlee March 26, 2018 at 2:27 pm #

REPLY ↩

How to copy code from the tutorial:

1. Click the copy button on the code example (top right of code box, second from the end). This will select all code in the box.
2. Copy the code to the clipboard (control-c on windows, command-c on mac, or right click and click copy).
3. Open your text editor.
4. Paste the code from the clip board.

This will preserve all white space.

Does that help?



Davy April 14, 2017 at 10:14 pm #

REPLY ↩

Hi, one beginner question. What do we get after training is completed in supervised learning, for classification problem ? Do we get weights? How do i use the trained model after that in field for real classification application lets say? I didn't get the concept what happens if

weights: how do I use the trained model after that in real, for real classification application lets say: I didn't get the concept what happens if training is completed. I tried this example: https://github.com/fchollet/keras/blob/master/examples/mnist_mlp.py and it printed me accuracy and loss of test data. Then what now?



Jason Brownlee April 15, 2017 at 9:35 am #

REPLY ↩

See this post on how to train a final model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>



Manikandan April 14, 2017 at 11:36 pm #

REPLY ↩

Wow... It's really great stuff man.... Thanks you....



Jason Brownlee April 15, 2017 at 9:36 am #

REPLY ↩

I'm glad to hear that.



Wes April 15, 2017 at 3:16 am #

REPLY ↩

As a complete beginner, it sounds so cool to predict the future. Then I saw all these model and complicated stuff, how do I even begin. Thank you for this. It is really great!



Jason Brownlee April 15, 2017 at 9:40 am #

REPLY ↩

You're very welcome.



Manjushree Aithal April 16, 2017 at 7:41 am #

REPLY ↩

Hello Jason,

I just started following your step by step tutorial for machine learning. In importing libraries step I followed each and every steps you specified, install all libraries via conda, but still I'm getting the following error.

Traceback (most recent call last):

File "C:/Users/dell/PycharmProjects/machine-learning/load_data.py", line 13, in

from sklearn.linear_model import LogisticRegression

File "C:/Users/dell/Anaconda2/lib/site-packages/sklearn/linear_model/__init__.py", line 15, in

from .least_angle import (Lars, LassoLars, lars_path, LarsCV, LassoLarsCV,

File "C:/Users/dell/Anaconda2/lib/site-packages/sklearn/linear_model/least_angle.py", line 24, in

from ..utils import arrayfuncs, as_float_array, check_X_y

ImportError: DLL load failed: Access is denied.

Can you please help me with this?

Thank You!



Jason Brownlee April 16, 2017 at 9:33 am #

REPLY ↩

I have not seen this error and I don't know about windows sorry.

It looks like you might not have admin permissions on your workstation.



Olah Data Semarang April 17, 2017 at 3:03 pm #

REPLY ↩

Tutorial DEAR Version 2.1

<https://www.youtube.com/watch?v=drd11htJJC0>

A Data Envelopment Analysis (Computer) Program. This page describes the computer program Tutorial DEAP Version 2.1 which was written by Tim Coelli.



Jason Brownlee April 18, 2017 at 8:30 am #

REPLY ↩

Thanks for sharing the link.



Federico Carmona April 18, 2017 at 4:41 am #

REPLY ↩

Good afternoon Dr. Jason could help me with the next problem. How could you modify the KNN algorithm to detect the most relevant variables?



Jason Brownlee April 18, 2017 at 8:34 am #

REPLY ↩

You can use feature importance scores from bagged trees or gradient boosting.

Consider using sklearn to calculate and plot feature importance.



Bharath April 18, 2017 at 10:09 pm #

REPLY ↩

Thank u...



Jason Brownlee April 19, 2017 at 7:52 am #

REPLY ↩

I'm glad the post helped.



Amal April 26, 2017 at 6:14 pm #

REPLY ↩

Hi Jason

Thanx for the great tutorial you provided.

I'm also new to MC and python. I tried to use my csv file as you used iris data set. Though it successfully loaded the dataset gives following error.

could not convert string to float: LipCornerDepressor

LipCornerDepressor is normal value such as 0.32145 in excel sheet taken from sql server

Here is the code without library files.

```
# Load dataset
```

```
url = "F:\FINAL YEAR PROJECT\Amila\FTdata.csv"
```

```
names = ['JawLower', 'BrowLower', 'BrowRaiser', 'LipCornerDepressor', 'LipRaiser', 'LipStretcher', 'Emotion_Id']
```

```
dataset = pandas.read_csv(url, names=names)
```

```
# shape
```

```
print(dataset.shape)
```

```
# class distribution
```

```
print(dataset.groupby('Emotion_Id').size())
```

```
# Split-out validation dataset
```

```
array = dataset.values
```

```
X = array[:,0:4]
```

```
Y = array[:,4]
```

```
validation_size = 0.20
```

```
seed = 7
```

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

```
# Test options and evaluation metric
seed = 7
scoring = 'accuracy'

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```



Jason Brownlee April 27, 2017 at 8:37 am #

REPLY ↩

This error might be specific to your data.

Consider double checking that your data is loaded as you expect. Maybe print some raw data or plots to confirm.



Chanaka April 27, 2017 at 6:31 am #

REPLY ↩

Thank you very much for the easy to follow tutorial.



Jason Brownlee April 27, 2017 at 8:48 am #

REPLY ↩

I'm glad you found it useful.



Sonali Deshmukh April 27, 2017 at 7:07 pm #

REPLY ↩

Hi, Jason

Your posts are really good.....

I'm very naive to Python and Machine Learning.

Can you please suggest good reads to get basic clear for machine learning.



Jason Brownlee April 28, 2017 at 7:38 am #

REPLY ↩

Thanks.

A good place to start for python machine learning is here:

<http://machinelearningmastery.com/start-here/#python>

I hope that helps.



Ianndo April 28, 2017 at 2:26 am #

REPLY ↩

Outstanding work on this. I am curious how to port out results that show which records were matched to what in the predictor, when I print(predictions) it does not show what records they are paired with. Thanks!



Jason Brownlee April 28, 2017 at 7:51 am #

REPLY ↩

Thanks!

The index can be used to align predictions with inputs. For example, the first prediction is for the first input, and so on.



NAVKIRAN KAUR April 29, 2017 at 4:28 pm #

REPLY ↩

when I am applying all the models and printing message it shows me the error that it cannot convert string to float. how to resolve this error. my data set is related to fake news ... title, text, label



Jason Brownlee April 30, 2017 at 5:27 am #

REPLY ↩

Ensure you have converted your text data to numerical values.



Shravan May 1, 2017 at 6:29 am #

REPLY ↩

Awesome tutorial on basics of machine learning using Python. Thank you Jason!



Jason Brownlee May 2, 2017 at 5:51 am #

REPLY ↩

Thanks Shravan.



Shravan May 1, 2017 at 6:36 am #

REPLY ↩

Am using Anaconda Python and I was writing all the commands/ program in the 'python' command line, am trying to find a way to save this program to a file? I have tried '%save', but it errored out, any thoughts?



Jason Brownlee May 2, 2017 at 5:51 am #

REPLY ↩

You can write your programs in a text file then run them on the command line as follows:

```
1 python file.py
```



Jason May 1, 2017 at 2:05 pm #

REPLY ↩

Thank you for the help and insight you provide. When I run the actual validation data through the algorithms, I get a different feel for which one may be the best fit.

Validation Test Accuracy:

LR.....0.80

LDA.....0.97

KNN....0.90

CART..0.87

NB.....0.83

SVM....0.93

My question is, should this influence my choice of algorithm?

Thank you again for providing such a wealth of information on your blog.



Jason Brownlee May 2, 2017 at 5:56 am #

REPLY ↩

Tes it should.

ML algorithms are stochastic and you need to evaluate them in such a way to take this into account.

This post might clarify what I mean:

<http://machinelearningmastery.com/randomness-in-machine-learning/>



rahman May 3, 2017 at 11:09 pm #

REPLY ↩

Split-out validation dataset

```
array = dataset.values
```

```
X = array[:,0:4]
```

```
Y = array[:,4]
```

from my dataset , When i give Y=array[:,1] Its working , but if give 2 or 3 or 4 instead of 1 it gives following error !!
But all columns have similar kind of data .

Traceback (most recent call last):

File "/alok/c-analyze/analyze.py", line 390, in

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 140, in cross_val_score
for train, test in cv_iter)

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 758, in __call__

```
while self.dispatch_one_batch(iterator):
```

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 608, in dispatch_one_batch
self._dispatch(tasks)

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 571, in _dispatch

```
job = self._backend.apply_async(batch, callback=cb)
```

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in apply_async

```
result = ImmediateResult(func)
```

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 326, in __init__

```
self.results = batch()
```

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__

```
return [func(*args, **kwargs) for func, args, kwargs in self.items]
```

File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 238, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)

File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 468, in fit

```
self._solve_svd(X, y)
```

File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 378, in _solve_svd

```
fac = 1. / (n_samples - n_classes)
```

ZeroDivisionError: float division by zero



Jason Brownlee May 4, 2017 at 8:08 am #

REPLY ↩

Perhaps take a closer look at your data.



rahman May 4, 2017 at 4:29 pm #

REPLY ↩

But the very similar in all the columns .



rahman May 4, 2017 at 4:37 pm #

REPLY ↩

I meant there is no much difference in data from each columns ! but still its working only for first column !! It gives the above error for any other column i choose .



rahman May 4, 2017 at 4:46 pm #

Have a look at the data :

```
index,1column,2 column,3column,....,8column
0,238,240,1103,409,1038,4,67,0
1,41,359,995,467,1317,8,71,0
2,102,616,1168,480,1206,7,59,0
3,0,34,994,181,1115,4,68,0
4,88,1419,1175,413,1060,8,71,0
5,826,10886,1316,6885,2086,263,119,0
6,88,472,1200,652,1047,7,64,0
7,0,322,957,533,1062,11,73,0
8,0,200,1170,421,1038,5,63,0
9,103,1439,1085,1638,1151,29,66,0
10,0,1422,1074,4832,1084,27,74,0
11,1828,754,11030,263845,1209,10,79,0
12,340,1644,11181,175099,4127,13,136,0
13,71,1018,1029,2480,1276,18,66,1
14,0,3077,1116,1696,1129,6,62,0
```

```
.....
.....
```

Total 105 data records

But the above error does not occur for 1 column , that is when Y = 1 column,
But the above same error happens when i choose any other column 2 , 3 or 4 .



hairo May 3, 2017 at 11:13 pm #

REPLY ↩

How to plot the graph for actual value against the predicted value here ?

How to save this plotted graphs and again view them back when required from terminal itself ?



Jason Brownlee May 4, 2017 at 8:08 am #

REPLY ↩

It would make for a dull graph as this is a classification problem.

You might be better of reviewing the confusion matrix of a set of predictions.



Sudarshan May 5, 2017 at 12:18 pm #

REPLY ↩

How this can be applied to predict the value if stastical dataset is given

Say i have given with past 10 years house price now i want to predict the value for house in next one year, two year

Can you help me out in this

I m amature in ML

Thank for this tutorial

It gives me a good kickstart to ML

I m waiting for your reply



Jason Brownlee May 6, 2017 at 7:30 am #

REPLY ↩

This is called a time series forecasting problem.

You can learn more about how to work through time series forecasting problems here:

<http://machinelearningmastery.com/start-here/#timeseries>



Sudarshan May 6, 2017 at 3:15 pm #

REPLY ↩

I getting trouble in doing that please help me out with any simple example

Example I have a dataset containing plumber work Say
attributes are

experience_level , date, rating, price/hour

I want to predict the price/hour for the next date base on experience level and average rating can you please help me regarding this.



Jason Brownlee May 7, 2017 at 5:34 am #

REPLY ↩

Sorry, I cannot write an example for you.



Bane May 8, 2017 at 4:30 am #

REPLY ↩

Great job with the tutorial, it was really helpful.

I want to ask, how can I use the techics above with a dataset that is not just one line with a few values, but a matrix NX3 with multiple values (measurements from an accelerometer). Is there a tutorial? How can I look up to it?



Jason Brownlee May 8, 2017 at 7:46 am #

REPLY ↩

Each feature would be a different input variable as in the example above.



Shud May 9, 2017 at 12:04 am #

REPLY ↩

Hey Jason,

I have built a linear regression model. y intercept is abnormally high (0.3 million) and adjusted r2 = 0.94. I would like to know what does high intercept mean?



Jason Brownlee May 9, 2017 at 7:45 am #

REPLY ↩

Think of the intercept as the bias term.

Many books have been written on linear regression and much is known about how to analyze these models effectively. I would recommend diving into the statistics literature.



MK May 11, 2017 at 12:19 am #

REPLY ↩

Excellent tutorial, i am moving from PHP to Python and taking baby steps. I used the Thonny IDE (<http://thonny.org/>) which is also very useful for python beginners.



Jason Brownlee May 11, 2017 at 8:33 am #

REPLY ↩

Thanks for sharing.



Tmoe May 14, 2017 at 4:31 am #

REPLY ↩

Thank you so much, Jason! I'm new to machine learning and python but found your tutorial extremely helpful and easy to follow – thank you for posting!



Jason Brownlee May 14, 2017 at 7:32 am #

REPLY ↩

Thanks Tmoe, I'm really glad to hear that!



melody12ab May 15, 2017 at 6:07 pm #

REPLY ↩

Thanks for all,now I am starting use ML!!!



Jason Brownlee May 16, 2017 at 8:39 am #

REPLY ↩

I'm glad to hear that!



smith May 15, 2017 at 9:36 pm #

REPLY ↩

Spot Check Algorithms

```
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
```

When i print models , this is the output :

```
[('LR', LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)), ('LDA', LinearDiscriminantAnalysis(n_components=None, priors=None, shrinkage=None,
solver='svd', store_covariance=False, tol=0.0001)), ('KNN', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
metric_params=None, n_jobs=1, n_neighbors=5, p=2,
weights='uniform'))]
```

What are these extra values inside LogisticRegression (...) and for all the other algorithms ?

How did they get appended ?



Jason Brownlee May 16, 2017 at 8:43 am #

REPLY ↩

You can learn about them in the sklearn API:

<http://scikit-learn.org/stable/modules/classes.html>



pasha May 15, 2017 at 9:45 pm #

REPLY ↩

When i print kfold :

```
KFold(n_splits=7, random_state=7, shuffle=False)
```

What is shuffle ? How did this value get added , as we had only done this :

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
```



Jason Brownlee May 16, 2017 at 8:44 am #

REPLY ↩

Whether or not to shuffle the dataset prior to splitting into folds.



pasha May 16, 2017 at 3:17 pm #

REPLY ↩

Now i understand , jason thanks for amazing tutorials . Just one suggestion along with the codes give a link for reference in detail about this topics !



Jason Brownlee May 17, 2017 at 8:24 am #

REPLY ↩

Great suggestion, thanks pasha.



sita May 15, 2017 at 9:48 pm #

REPLY ↩

Hello jason

This is an amazing blog , Thank you for all the posts .

`cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)`

Whats scoring here ? can you explain in detail " `model_selection.cross_val_score` " this line please .



Jason Brownlee May 16, 2017 at 8:45 am #

REPLY ↩

Thanks sita.

Learn more here:

[http://scikit-](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score)

[learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score)



rahman May 15, 2017 at 10:27 pm #

REPLY ↩

Please help me with this error Jason ,

ERROR :

Traceback (most recent call last):

File "rahman/c-analyze/analyze.py", line 390, in

`cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)`

File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 140, in `cross_val_score`
for train, test in `cv_iter`)

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 758, in `__call__`

while `self.dispatch_one_batch(iterator)`:

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 608, in `dispatch_one_batch`

`self._dispatch(tasks)`

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 571, in `_dispatch`

`job = self._backend.apply_async(batch, callback=cb)`

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in `apply_async`

`result = ImmediateResult(func)`

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 326, in `__init__`

`self.results = batch()`

File "/usr/lib64/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in `__call__`

return `[func(*args, **kwargs) for func, args, kwargs in self.items]`

File "/usr/lib64/python2.7/site-packages/sklearn/model_selection/_validation.py", line 238, in `_fit_and_score`

`estimator.fit(X_train, y_train, **fit_params)`

File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 468, in `fit`

`self._solve_svd(X, y)`

File "/usr/lib64/python2.7/site-packages/sklearn/discriminant_analysis.py", line 378, in `_solve_svd`

`fac = 1. / (n_samples - n_classes)`

ZeroDivisionError: float division by zero

Split-out validation dataset

My code :

`array = dataset.values`

`X = array[:,0:4]`

if `field == "rh"`: #No error if i select this col

`Y = array[:,0]`

elif field == "rm": #gives the above error

Y = array[:,1]

elif field == "wh": #gives the above error

Y = array[:,2]

elif field == "wm": #gives the above error

Y = array[:,3]

Have a look at the data :

index,1column,2 column,3column,....,8column

0,238,240,1103,409,1038,4,67,0

1,41,359,995,467,1317,8,71,0

2,102,616,1168,480,1206,7,59,0

3,0,34,994,181,1115,4,68,0

4,88,1419,1175,413,1060,8,71,0

5,826,10886,1316,6885,2086,263,119,0

6,88,472,1200,652,1047,7,64,0

7,0,322,957,533,1062,11,73,0

8,0,200,1170,421,1038,5,63,0

9,103,1439,1085,1638,1151,29,66,0

10,0,1422,1074,4832,1084,27,74,0

11,1828,754,11030,263845,1209,10,79,0

12,340,1644,11181,175099,4127,13,136,0

13,71,1018,1029,2480,1276,18,66,1

14,0,3077,1116,1696,1129,6,62,0

.....

.....

Total 105 data records

But the above error does not occur for 1 column , that is when Y = 1 column,

But the above same error happens when i choose any other column 2 , 3 or 4 .



Jason Brownlee May 16, 2017 at 8:45 am #

REPLY ↩

Perhaps try scaling your data?

Perhaps try another algorithm?



suma May 16, 2017 at 12:05 am #

REPLY ↩

fac = 1. / (n_samples - n_classes)

ZeroDivisionError: float division by zero

What is this error : fac = 1. / (n_samples - n_classes) ?

Where is n_samples and n_classes used ?

What may be the possible reason for this error ?



bob May 22, 2017 at 6:46 pm #

REPLY ↩

thank you Dr Jason it is really very helpfully. 😊



Jason Brownlee May 23, 2017 at 7:50 am #

REPLY ↩

You're welcome bob, I'm glad to hear that!



Krithika May 24, 2017 at 12:24 am #

REPLY ↩

Hi Jason

Great starting tutorial to get the whole picture. Thank you:)

I am a newbie to machine learning. Could you please tell why you have specifically chosen these 6 models?



Jason Brownlee May 24, 2017 at 4:57 am #

REPLY ↩

No specific reason, just a demonstration of spot checking a suite of methods on the problem.



Ram Gour May 25, 2017 at 8:24 pm #

REPLY ↩

Hi Jason, I am new to Python, but found this blog really helpful. I tried executing the code and it return all the result as mention above by you, except few graph.

The scatter matrix graph and the evaluation on 6 algorithm did not open on my machine but its showing result on my colleague machine. I checked all the version and its higher or same as you mentioned in blog.

Can you help if this issue can be resolved on my machine?



Jason Brownlee June 2, 2017 at 11:44 am #

REPLY ↩

Perhaps check the configuration of matplotlib and ensure you can create simple graphs on your machine?



sridhar May 25, 2017 at 8:50 pm #

REPLY ↩

Great tutorial.

How do I approach when the data set is not of any classification type and the number of attributes or just 2 – 1 is input and the other is output

say I have number of processes as input and cpu usage as output..

data set looks like [10, 5] [15, 7] etc...



Jason Brownlee June 2, 2017 at 11:45 am #

REPLY ↩

If the output is real-valued, it would be a regression problem. You would need to use a loss function like MSE.



pierre May 27, 2017 at 9:45 pm #

REPLY ↩

Many thanks for this — I already got a lot out of this. I feel like a monkey though because I was neither familiar enough with python nor had any clue of ML back alleys yesterday. Today I can see plots on my screen and even if I have no clue what I'm looking at, this is where I wanted to be, so thanks!

A few minor suggestions to make this perhaps even more dummy-proof:

– I'm on Mac and I used python3 because python2 is weirdly set up out of the box and you can't update easily the libraries needed. I understand you link, rightfully to external installation instructions, so just to say, this stuff works in python3 if you needed further testimony.

– when drawing plots, I started freaking out because the terminal became unresponsive. So if you just made an (unessential) suggestion to run `plt.ion()` first, linking to, for example: https://matplotlib.org/faq/usage_faq.html#what-is-interactive-mode, it might help dummies like me to not give up too easily. (BTW I find your use command line philosophy and don't let toolsets get in the way a great one indeed!)

– There seems to be some 'hack' involved when defining the dataset, suppose there are no headers and so on... how do you get to load your dataset with an insightful name vector in the first place (you don't...) So just a hint of clarification would help here feeling we can trust that we do the right thing in this case because the data is well understood (I mean, this is not really a big deal eh it's all par for the course but if I didn't have similar experience in R I'd feel completely lost I think).

I was a bit puzzled by the following sentence in 3.3:

"We can see that all of the numerical values have the same scale (centimeters) and similar ranges between 0 and 8 centimeters."

Well, just looking at the table, I actually can't see any of this. There is in fact really nothing telling this to us in the snippet, right? The sentence is a comment based on prior understanding of the dataset. Maybe this could be clarified so clueless readers don't agonise over whether they are missing some magical power of insight.

– Overall, I could run this and to some extent adapt it quickly to a different dataset until it became relevant what the data was like. I'm stumbling on the data manipulation for 5.1. I suppose it is both because I don't know python structures and also because I have no clue what is being done in the selection step.

I think in answer to a previous comment you link to doc for the relevant selection function, perhaps it would still be useful to have an extra, 'for dummies', detailed explanation of

```
X = array[:,0:4]
Y = array[:,4]
```

in the context of the iris dataset. This is what I have to figure out, I think, in order to apply it to say, a 11 column dataset and it would be useful to know what I'm trying to do.

The rest of the difficulties I have are with regards to interpretation of the output and it is fair to say this is outside of the scope of your tutorial which puts dummies like me in a very good position to try to understand while being able to fiddle with a bit of code. All the above comments are extremely minor and really about polishing the readability for ultimate noobs, they are not really important and your tutorial is a great and efficient resource.

Thanks again!
Pierre



Jason Brownlee June 2, 2017 at 12:04 pm #

REPLY ↩

Wonderful feedback pierre, thank you so much!



Shaksham Kapoor June 6, 2017 at 4:18 am #

REPLY ↩

I'm not able to figure out , what errors does the confusion matrix represents ? and what does each column(precision, recall, f1-score, support) in the classification report signifies ?

And last but not the least thanks a lot Sir for this easy to use and wonderful tutorial. Even words are not enough to express my gratitude, you have made a daunting task for every ML Enthusiast a hell lot easier !!!



Jason Brownlee June 6, 2017 at 10:07 am #

REPLY ↩

You can learn more about the confusion matrix here:

<http://machinelearningmastery.com/confusion-matrix-machine-learning/>



Shaksham Kapoor June 7, 2017 at 3:39 am #

REPLY ↩

Thanks a lot Sir. Please suggest some data-sets from UCL repository on which I can practice some small projects...



Jason Brownlee June 7, 2017 at 7:26 am #

REPLY ↩

See here:

<http://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/>



Shaksham Kapoor June 7, 2017 at 6:48 pm #

How do you classify problem into different categories example : Iris dataset was a classification problem and pima-indian-diabetes ,a binary problem. How can we figure out which problem belong to which category and which model to apply on that problem?



Jason Brownlee June 8, 2017 at 7:40 am #

By careful evaluation of the output variable.



Brian June 6, 2017 at 11:11 pm #

REPLY ↩

Is this machine learning? what does the machine learn in this example? This is just plain Statistics, used in a weird way...



Jason Brownlee June 7, 2017 at 7:14 am #

REPLY ↩

Yes, it is.

Nominally, statistics is about understanding the data, machine learning about making predictions at the cost of understanding.



Raj June 9, 2017 at 2:22 am #

REPLY ↩

your question can be answered like this...

consider the formula for area of triangle $\frac{1}{2} \times \text{base} \times \text{height}$. When you learn this formula, you understand it and apply it many times for different triangles. BUT you did not learn anything ABOUT the formula itself. . for instance, how many people care that the formula has 2 variables(base and height) and that there is no CONSTANT(like PI) in the formula and many such things about the formula itself? Applying the formula does not teach anything about the nature of the formula itself

A lot of program execution in computers happen much the same way...data is a thing to be modified, applied or used, but not necessarily understood. When you introduce some techniques to understand data, then necessarily the computer or the 'Machine' 'learns' that there are characteristics about that data, and that at the least, there exists some relationship amongst data in their dataset. This learning is not explicitly programmed rather inferred, although confusingly, the algorithms themselves are explicitly programmed to infer the meaning of the dataset. The learning is then transferred to the end cycle of making prediction based on the gained understanding of data.

but like you pointed out, it is still statistics and all it's domain techniques, but as a statistician do you not 'learn' more about data than merely use it, unlike your counterparts who see data more as a commodity to be consumed? Because most computer systems do the latter(consumption) rather than the former(data understanding), a system that understands data(with prediction used as a proof of learning) can be called 'Machine Learning'.



Alex June 7, 2017 at 6:04 am #

REPLY ↩

Thanks for good tutorial Jason.

Only issue I encountered is following error while cross validation score calculation for model KNeighborsClassifier() :

AttributeError: 'NoneType' object has no attribute 'issparse'

Is somebody got same error? How it can be solved?

I have installed following versions of toos:

Python: 2.7.13 [Anaconda custom (64-bit)] (default, Dec 19 2016, 13:29:36) [MSC v.1500 64 bit (AMD64)]

scipy: 0.19.0

numpy: 1.12.1

matplotlib: 2.0.0

pandas: 0.19.2

sklearn: 0.18.1

Thanks,
Alex



Jason Brownlee June 7, 2017 at 7:27 am #

REPLY ↩

Ouch, sorry I have not seen this issue. Perhaps search on stackoverflow?



thanda June 8, 2017 at 6:31 pm #

REPLY ↩

Hi, Jason!

How can i get the xgboost algorithm in pseudo code or in code?



Jason Brownlee June 9, 2017 at 6:21 am #

REPLY ↩

You can read the code here:

<https://github.com/dmlc/xgboost>

I expect it is deeply confusing to read.

For an overview of gradient boosting, see this post:

<http://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>



Shaksham Kapoor June 9, 2017 at 1:14 am #

REPLY ↩

Sir,I've been working on bank_note authentication dataset and after applying the above procedure carefully the results were 100% accuracy(both on trained and validation dataset) using SVM and KNN models. Is 100% accuracy possible or have I done something wrong ?



Jason Brownlee June 9, 2017 at 6:27 am #

REPLY ↩

That sounds great.

If I were to get surprising results, I would be skeptical of my code/models.

Work hard to ensure your system is not fooling you. Challenge surprising results.



Shaksham Kapoor June 9, 2017 at 3:10 pm #

REPLY ↩

Sir, I've considered various other aspects like f1-score, recall, support ; but in each case the result is same 100%. How can I make sure that my system is not fooling me ? What other procedure can I apply to check the accuracy of my dataset ?



Jason Brownlee June 10, 2017 at 8:13 am #

REPLY ↩

Get more data and see if the model can make accurate predictions.



Rejeesh R June 9, 2017 at 7:27 pm #

REPLY ↩

Hi, Jason!

I am new to python as well ML. so I am getting the below error while running your code, please help me to code bring-up

File "sample1.py", line 73, in

predictions = knn.predict(X_validation)

File "/usr/local/lib/python2.7/dist-packages/sklearn/neighbors/classification.py", line 143, in predict

X = check_array(X, accept_sparse='csr')

File "/usr/local/lib/python2.7/dist-packages/sklearn/neighbors/classification.py", line 143, in predict

_assert_all_finite(array)

File "/usr/local/lib/python2.7/dist-packages/sklearn/neighbors/classification.py", line 143, in predict

" or a value too large for %r." % X.dtype)

ValueError: Input contains NaN, infinity or a value too large for dtype('float64').

and my config

Python: 2.7.6 (default, Oct 26 2016, 20:30:10)

Python: 2.7.6 (default, Oct 26 2016, 20:30:19)
[GCC 4.8.4]
scipy: 0.13.3
numpy: 1.8.2
matplotlib: 1.3.1
pandas: 0.13.1
sklearn: 0.18.1
running in Ubuntu Terminal.



Jason Brownlee June 10, 2017 at 8:20 am #

REPLY ↩

You may have a NaN value in your dataset. Check your data file.



Sats S June 10, 2017 at 5:27 am #

REPLY ↩

Hello. This is really an amazing tutorial. I got down to everything but when selecting the best model i hit a snag. Can you help out?

Traceback (most recent call last):

```
File "/Users/sahityasehgal/Desktop/py/machinetest.py", line 77, in
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/model_selection/_validation.py", line 140, in cross_val_score
for train, test in cv_iter)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/parallel.py", line 758, in __call__
while self.dispatch_one_batch(iterator):
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/parallel.py", line 608, in dispatch_one_batch
self._dispatch(tasks)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/parallel.py", line 571, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in
apply_async
result = ImmediateResult(func)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 326, in __init__
self.results = batch()
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/model_selection/_validation.py", line 238, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/linear_model/logistic.py", line 1173, in fit
order="C")
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/utils/validation.py", line 526, in check_X_y
y = column_or_1d(y, warn=True)
File "/Users/sahityasehgal/Library/Python/2.7/lib/python/site-packages/sklearn/utils/validation.py", line 562, in column_or_1d
raise ValueError("bad input shape {0}".format(shape))
ValueError: bad input shape (94, 4)
```



Jason Brownlee June 10, 2017 at 8:28 am #

REPLY ↩

Ouch. Are you able to confirm that you copied all of the code exactly?

Also, are you able to confirm that your sklearn is up to date?



Sats S June 10, 2017 at 11:10 am #

REPLY ↩

Yes i coped the code exactly as on the site. sklearn: 0.18.1 thoughts?



Jason Brownlee June 11, 2017 at 8:20 am #

REPLY ↩

I'm not sure but I expect it has something to do with your environment.

This tutorial may help with your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Rene June 11, 2017 at 1:25 am #

REPLY ↩

Very insightful Jason, thank you for the post!

I was wondering if the models can be saved to/loaded from file, to avoid re-training a model each time we wish to make a prediction.

Thanks,

Rene



Jason Brownlee June 11, 2017 at 8:26 am #

REPLY ↩

Yes, see this post:

<http://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/>



Richard Bruning June 12, 2017 at 11:42 am #

REPLY ↩

Mr. Brownlee,

This is, by far, is the most effective applied technology tutorial I have utilized.

You get right to the point and still have readers actually working with python, python libraries, IDE options, and of course machine learning. I am an electromechanical engineer with embedded C experience. Until now, I have been bogged down trying to traipse through python wizards' idiosyncratic coding styles and verbose machine learning theory knowing there exists a friendlier path.

Thank you for showing me the way!

Rich



Jason Brownlee June 13, 2017 at 8:13 am #

REPLY ↩

Thanks Rich, you made my day! I'm glad it helped.



Prayer Vats June 13, 2017 at 7:21 pm #

REPLY ↩

This was very informative....Thank You !

Actually I was working on a project on twitter analysis using python where I am extracting user interests through their tweets. I was thinking of using naive bayes classifier in textblob python library for training classifier with different type of pre-labeled tweets or different categories like politics,sports etc.

My only concern is that will it be accurate as I tried passing like 10 tweets in training set and based on that I tried classifying my test set. I am getting some false cases and accuracy is around 85.



Jason Brownlee June 14, 2017 at 8:44 am #

REPLY ↩

Good question, I'd suggest try it and see.



Kush Singh Kushwaha June 14, 2017 at 4:14 am #

REPLY ↩

Hi Jason,

This was great example. I was looking for something similar on internet all this time,glad I found this link. I wanted to compile a ML code end-to-end and see my basic infra is ready to start with the actual course work. As you said, from here we can learn more about each

algorithm in detail. It would be great if you can start a Youtube channel and upload some easy to learn videos as well related to ML, Deep learning and Neural Networks.

Regards,
Kush Singh



Jason Brownlee June 14, 2017 at 8:51 am #

REPLY ↩

Thanks.

Take a look at the rest of my blog and my books. I am dedicated to this mission.



Shaksham Kapoor June 14, 2017 at 4:34 am #

REPLY ↩

I've been working on a dataset which contains [Male,Female,Infant] as entries in first column rest all columns are integers. How can I replace [Male,Female,Infant] with a similar notation like [0,1,2] or something like that ? What is the most efficient way to do it ?



Jason Brownlee June 14, 2017 at 8:51 am #

REPLY ↩

Excellent question.

Use a LabelEncoder:

<http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

I'm sure I have tutorials on this on my blog, try the blog search.



Dev June 14, 2017 at 12:52 pm #

REPLY ↩

Sir, while loading dataset we have given the URI but what if we already have one and wants to load it ?



Jason Brownlee June 15, 2017 at 8:42 am #

REPLY ↩

Change the URL to a filename and path.



Vincent June 18, 2017 at 2:26 am #

REPLY ↩

Hi,

Nice tutorial, thanks!

Just a little precision if someone encounter the same issue than me:

if you get the error "This application failed to start because it could not find or load the Qt platform plugin "windows" in ""." when you are trying to see your data visualizations, it's maybe (like in my case) because you are using PySide rather than PyQt. In that case, add these lines before the "import matplotlib.pyplot as plt":

```
import matplotlib
matplotlib.use('Qt4Agg')
matplotlib.rcParams['backend.qt4']='PySide'
```

Hope this will help



Jason Brownlee June 18, 2017 at 6:33 am #

REPLY ↩

Thanks for the tip Vincent.



Danielle June 25, 2017 at 5:43 am #

REPLY ↩



Fantastic tutorial! Running today I noticed two changes from the tutorial above (undoubtably because time has passed since it was created). New users might find the following observations useful:

#1 – Future Warning

Ran on OS X, Python 3.6.1, in a jupyter notebook, anaconda 4.4.0 installed:

```
scipy: 0.19.0
numpy: 1.12.1
matplotlib: 2.0.2
pandas: 0.20.1
sklearn: 0.18.1
```

I replaced this line in the #Load libraries code block:
from pandas.tools.plotting import scatter_matrix

With this:
from pandas.plotting import scatter_matrix

...because a FutureWarning popped up:
/Users/xxx/anaconda/lib/python3.6/site-packages/ipykernel_launcher.py:2: FutureWarning: 'pandas.tools.plotting.scatter_matrix' is deprecated, import 'pandas.plotting.scatter_matrix' instead.

Note: it does run perfectly even without this fix, this may be more of an issue in the future

#2 – SVM wins!

In the build models section, the results were:

```
LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.966667 (0.040825)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)
```

... which means SVM was better here. I added the following code block based on the KNN one:

```
# Make predictions on validation dataset
svm = SVC()
svm.fit(X_train, Y_train)
predictions = svm.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

which gets these results:

```
0.933333333333
[[ 7 0 0]
 [ 0 10 2]
 [ 0 0 11]]
precision recall f1-score support
Iris-setosa 1.00 1.00 1.00 7
Iris-versicolor 1.00 0.83 0.91 12
Iris-virginica 0.85 1.00 0.92 11

avg / total 0.94 0.93 0.93 30
```

I did also run the unmodified KNN block – # Make predictions on validation dataset – and got the exact results that were in the tutorial.

Excellent tutorial, very clear, and easy to modify 🙌



Jason Brownlee June 26, 2017 at 6:06 am #

REPLY ↩

Thanks for sharing Danielle.



mr. disapointed June 26, 2017 at 10:06 pm #

REPLY ↩

So this intro shows how to set everything up but not the actual interesting bit how to use it?



Jason Brownlee June 27, 2017 at 8:29 am #

REPLY ↩

What do you mean exactly? Putting the model into production? See here:
<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>



Aditya June 28, 2017 at 4:48 pm #

REPLY ↩

Excellent tutorial sir, I love your tutorials and I am starting with deep learning with keras.
I would love if you could provide a tutorial for sequence to sequence model using keras and a relevant dataset.
Also I would be obliged if you could point me in some direction towards names entity recognition using seq2seq



Jason Brownlee June 29, 2017 at 6:29 am #

REPLY ↩

I have one here:
<http://machinelearningmastery.com/learn-add-numbers-seq2seq-recurrent-neural-networks/>



RATNA June 30, 2017 at 4:19 am #

REPLY ↩

Hi Jason,

Awesome tutorial. I am working on PIMA dataset and while using the following command
head
print(dataset.head(20))
I am getting NAN. HEPL ME.



Jason Brownlee June 30, 2017 at 8:18 am #

REPLY ↩

Confirm you downloaded the dataset and that the file contains CSV data with nothing extra or corrupted.



RATNA June 30, 2017 at 4:14 pm #

REPLY ↩

Hi Jason,

I downloaded the dataset from UCI which is a CSV file but still I get NAN.

Load dataset url = "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data"

Thanks..



Jason Brownlee July 1, 2017 at 6:27 am #

REPLY ↩

Sorry, I do not see how this could be. Perhaps there is an issue with your environment?



Deepak July 2, 2017 at 1:50 am #

REPLY ↩

Hello Jason,

Thank you for a great tutorial.

I have noticed something , which I would like to share with you.

I have tried with random_state = 4

"X_train,X_validation,Y_train,Y_validation = model_selection.train_test_split(X,Y, test_size = 0.2, random_state = 4)"

and surprisingly now "LDA" has the best accuracy.

LR: 0.966667 (0.040825)
LDA: 0.991667 (0.025000)
KNN: 0.975000 (0.038188)
CART: 0.958333 (0.055902)
NB: 0.950000 (0.055277)
SVM: 0.983333 (0.033333)

Any thoughts on this?



Jason Brownlee July 2, 2017 at 6:33 am #

REPLY ↩

Machine learning algorithms are stochastic:

<http://machinelearningmastery.com/randomness-in-machine-learning/>



Rui July 3, 2017 at 12:31 pm #

REPLY ↩

Hi Jason,

Thanks for your great example, this is really helpful, this end-to-end project is the best way to learn ML, much better than text-book which they only focus on the separate concepts, not the whole forest, will you please do more example like this and explain in detail next time?

Thanks,

Rui



Jason Brownlee July 6, 2017 at 9:57 am #

REPLY ↩

Thanks.



Vaibhav July 4, 2017 at 4:33 pm #

REPLY ↩

`__init__()` got an unexpected keyword argument 'n_splits'

I am getting this error while running the code upto "print(msg)" command.

Can you please help me removing it.



Jason Brownlee July 6, 2017 at 10:12 am #

REPLY ↩

Update your version of sklearn to 0.18 or higher.



Fahad Ahmed July 5, 2017 at 12:31 am #

REPLY ↩

This is beautiful tutorial for the starters..

I am a lover of machine learning and want to do some projects and research on it.

I would really need your help and guideline time to time.

Regards,

Fahad



Jason Brownlee July 6, 2017 at 10:19 am #

REPLY ↩

Thanks.



Neal Valiant July 12, 2017 at 9:08 am #

REPLY ↩



Hi Jason,

Love the article. gave me a good start of understanding machine learning. One thing i would like to ask is what is the predicted outcome? Is it which type or "class" of flower that will happen next? i assume switching things up I could use this same outline as a way of getting a prediction on the other columns involved?



Jason Brownlee July 12, 2017 at 9:55 am #

REPLY ↩

Yes, the prediction is a number that maps to a specific class of flower (string).

Correct, from the class and other measures you could predict width or something.



Neal July 13, 2017 at 3:50 am #

REPLY ↩

Hi again Jason,

Diving deeper into this tutorial and analyzing more I find something that peaked an interest maybe you can shed light on. based off the seed of 7 you get a higher accuracy percentage on the KNN algorithm after using kfold, but when showing the information for the LDA algorithm, it has a higher percentage in accuracy_score after predicting on it. what could this mean?



Jason Brownlee July 13, 2017 at 9:59 am #

REPLY ↩

Machine learning algorithms are stochastic.

It is important to develop a robust estimate of the performance of machine learning models on unseen data using repeats. See this post:

<http://machinelearningmastery.com/evaluate-skill-deep-learning-models/>



Neal July 13, 2017 at 11:22 am #

Another great read Jason. This whole site is full of great pieces and it gives me a good answer on my question. I want to thank you for your time and effort into making such a great place for all this knowledge.



Jason Brownlee July 13, 2017 at 4:54 pm #

Thanks, I'm glad it helps Neal. Stick with it!



Thomas July 14, 2017 at 8:10 pm #

REPLY ↩

Hello Jason,

At the beginning of your tutorial you write: "If you are a machine learning beginner and looking to finally get started using Python, this tutorial was designed for you."

No offense but in this regards, your tutorial is not doing a very good job.

You don't really go in detail so that we can understand what is been done and why. The explanations are rather weak.

Wrong expectations set i believe.

Cheers,

Thomas



Jason Brownlee July 15, 2017 at 9:43 am #

REPLY ↩

It is a starting point, not a panacea.

Sorry that it's not a good fit for you.



Mariah July 15, 2017 at 7:11 am #

REPLY ↩

Hi Jason! I am trying to adapt this for a purely binary dataset, however I'm running into this problem:

```
# evaluate each model in turn
results = []
name = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv = kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s:%f(%f)"%(name, cv_results.mean(), cv_results.std())
    print(msg)
```

I get the error:

```
raise ValueError("Unknown label type: %r" % y_type)
```

ValueError: Unknown label type: 'unknown'

Am I missing something, any help would be great!



Mariah July 15, 2017 at 7:12 am #

REPLY ↩

All necessary indentations are correct, it just pasted incorrectly



Jason Brownlee July 15, 2017 at 9:46 am #

REPLY ↩

You can wrap pasted code in pre tags.



Jason Brownlee July 15, 2017 at 9:46 am #

REPLY ↩

Sorry, the fault is not obvious to me.



Daniel September 12, 2017 at 1:14 am #

REPLY ↩

Hello Mariah,

Did you ever get a solution to this problem?

Jason..great guide here..THANKS!



Sreeram July 16, 2017 at 10:09 pm #

REPLY ↩

Hi. What should i do to make predictions based on my own test set.? Say i need to predict category of flower with data [5.2, 1.8, 1.6, 0.2]. ie i want to change my X_test to that array. And the prediction should be like "setosa".

What changes should i do.? I tried giving that value directly to predict(). But it crashes.



Jason Brownlee July 17, 2017 at 8:47 am #

REPLY ↩

Correct.

Fit the model on all available data. This is called creating a final model:

<http://machinelearningmastery.com/train-final-machine-learning-model/>

Then make your prediction on new data where you do not know the answer/outcome.

Does that help?



Sreeram July 18, 2017 at 2:35 am #

REPLY ↩

Yes it helped. Can u show an example code for the same.?



Jason Brownlee July 18, 2017 at 8:46 am #

REPLY ↩

Sure:

```
1 # train on all data
2 model = ...
3 # make prediction on new 1D instance
4 result = model.predict(newX)
```



Joe July 18, 2017 at 7:49 am #

REPLY ↩

Hi Jason, i'm perú and i have to script write in Mac

#Configurar para la red neural

fechantinicio = '1970-01-01'

fechantfinal = '1974-12-31'

capasinicio = TodasEstaciones.ix[fechantinicio:fechantfinal].as_matrix()[:[0,2,5]]

capasalida = TodasEstaciones.ix[fechantinicio:fechantfinal].as_matrix()[:[,1]]

#Construimos la Red Neural

from sknn.mlp import Regressor, Layer

neurones = 8

tasaaprendizaje = 0.0001

numiteraciones = 7000

#Definition of the training for the neural network

redneural = Regressor(

layers=[

Layer("Explin", units=neurones),

Layer("Explin", units=neurones), Layer("Linear")],

learning_rate=tasaaprendizaje,

n_iter=numiteraciones)

redneural.fit(capasinicio, capasalida)

#Get the prediction for the train set

valortest = ([])

for i in range(capasinicio.shape[0]):

prediccion = redneural.predict(np.array([capasinicio[i,:].tolist()]))

valortest.append(prediccion[0][0])

and then run...

ModuleNotFoundError Traceback (most recent call last)

in ()

1 #Construimos la Red Neural

2

—> 3 from sknn.mlp import Regressor, Layer

4

5

ModuleNotFoundError: No module named 'sknn'

i have install python in window 7 and i changed the script so:

#construimos la red neural

import numpy as np

from sklearn.neural_network import MLPRegressor

#definicion del entrenamiento para el trabajo de la red neural

redneural = MLPRegressor(

hidden_layer_sizes=(100), activation='relu', solver='adam', alpha=0.001, batch_size='auto'

```
hidden_layer_sizes=(100,)), activation='relu', solver='adam', alpha=0.001, batch_size='auto',
learning_rate='constant', learning_rate_init=0.01, power_t=0.5, max_iter=1000, shuffle=True,
random_state=0, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True,
early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08)
```

redneural.fit(capasinicio,capasalida) and then shift + enter the run never end.

Thanks for your time.



Jason Brownlee July 18, 2017 at 8:49 am #

REPLY ↩

Consider posting to stackoverflow.



Angel July 18, 2017 at 6:06 pm #

REPLY ↩

Hello Jason, this is a fantastic tutorial! I am using this as a template to experiment with a dataset that has 0 or 1 as a value for each attribute and keep running into this error:

```
# Load libraries
import numpy
from matplotlib import pyplot
from pandas import read_csv
from pandas import set_option
from pandas.tools.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier
# Load Dataset
filename = 'ML.csv'
names = ['Cities', 'Entertainment', 'RegionalFood', 'WestMiss', 'NFLTeam', 'Coastal', 'WarmWinter', 'SuperBowl', 'Manufacturing']
data = read_csv(filename, names=names)
print(data.shape)
# types
set_option('display.max_rows', 500)
print(data.dtypes)
# head
set_option('display.width', 100)
print(data.head(20))
# descriptions, change precision to 3 places
set_option('precision', 3)
print(data.describe())
# class distribution
print(data.groupby('Cities').size())
# histograms
data.hist(sharex=False, sharey=False, xlabelsize=1, ylabelsize=1)
pyplot.show()
# correlation matrix
fig = pyplot.figure()
```

```

fig = pyplot.figure()
ax = fig.add_subplot(111)
cax = ax.matshow(data.corr(), vmin=-1, vmax=1, interpolation='none')
fig.colorbar(cax)
pyplot.show()
# Split-out validation dataset
array = data.values
X = array[:,1:8]
Y = array[:,8]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = train_test_split(X, Y,
test_size=validation_size, random_state=seed)
# Test options and evaluation metric
num_folds = 3
seed = 7
scoring = 'accuracy'
# Spot-Check Algorithms
models = []
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
results = []
names = []
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = KFold(n_splits=3, random_state=seed)
    cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

```

I get the following error:

File "C:\Users\Giselle\Anaconda3\lib\site-packages\sklearn\utils\multiclass.py", line 172, in check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)

ValueError: Unknown label type: 'unknown'

runfile('C:/Users/Giselle/.spyder-py3/temp.py', wdir='C:/Users/Giselle/.spyder-py3')



Jason Brownlee July 19, 2017 at 8:22 am #

REPLY ↩

Check that you are loading your data correctly.



machine learning guy July 18, 2017 at 9:15 pm #

REPLY ↩

hey jason.

awesome detailed blog man....i always love your method for explanation ..so clean and easy. Great ... i start machine learning with r but now doing with python too.

Regards

Kuldeep



Jason Brownlee July 19, 2017 at 8:23 am #

REPLY ↩

Thanks.



Aayush A July 18, 2017 at 9:17 pm #

REPLY ↩

Hey Jason,

Your sample code is amazing to get started with ML.

When I tried to run the code myself I get an

Can you please help me rectify this?



Jason Brownlee July 19, 2017 at 8:23 am #

REPLY ↩

What is the problem?



Marco Roque July 19, 2017 at 7:01 am #

REPLY ↩

Jason

Thanks for your help !!!! The Blog is super useful ... do you have another place that you recommend to learn more about the topic
Thanks !!!!

Best

Marco



Jason Brownlee July 19, 2017 at 8:31 am #

REPLY ↩

Thanks.

Yes, search "resources" on the blog.



Yug July 20, 2017 at 2:59 am #

REPLY ↩

Hi Jason,

Great tutorial!! very helpful!

I am getting an error executing below piece of code, can you help?

```
# evaluate each model in turn
```

```
results = []
```

```
names = []
```

```
for name, model in models:
```

```
    kfold = ms.KFold(n_splits=10, random_state=seed)
```

```
    cv_results = ms.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

```
    results.append(cv_results)
```

```
    names.append(name)
```

```
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
```

```
    print(msg)
```

Error that I am getting:

TypeError: get_params() missing 1 required positional argument: 'self'



Jason Brownlee July 20, 2017 at 6:22 am #

REPLY ↩

Sorry, I have not seen that error before. Perhaps confirm that your environment is installed correctly?

Also confirm that you have all of the code without extra spaces?



Yug July 20, 2017 at 8:02 am #

REPLY ↩

Yeah, environment is installed correctly. I made sure that there are no extra spaces in the code. It is still erroring out.



Jason Brownlee July 21, 2017 at 9:23 am #

REPLY ↩

Sorry, I'm running out of ideas.



Aawesh July 21, 2017 at 8:40 am #

REPLY ↩

Great tutorial. Loved it. What's next?



Jason Brownlee July 21, 2017 at 9:37 am #

REPLY ↩

See here:

<http://machinelearningmastery.com/start-here/#python>

And for the higher-level goals (e.g. build a portfolio):

<http://machinelearningmastery.com/start-here/#getstarted>



Chandana July 21, 2017 at 8:54 am #

REPLY ↩

I get the following results when the test is run against each model.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.966667 (0.040825)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Looks like SVN is the best and not KNN, what is the reason for this?



Jason Brownlee July 21, 2017 at 9:37 am #

REPLY ↩

Machine learning algorithms are stochastic:

<http://machinelearningmastery.com/randomness-in-machine-learning/>



samkelo jiyane July 21, 2017 at 4:24 pm #

REPLY ↩

Hi Jason, have started to learn Machine learning basics using Keras (with TF/Theano as backend). I am going through examples on this site and other resources with the ultimate goal of implementing Document reading/interpretation on constrained data set, e.g bank statements, proof of residence, standard supporting document etc.

Any pointers ?



Jason Brownlee July 22, 2017 at 8:30 am #

REPLY ↩

Great!

Yes, start here:

<http://machinelearningmastery.com/start-here/#getstarted>



Asad Ali July 23, 2017 at 1:04 pm #

REPLY ↩

Thank you Jason for this simple tutorial for beginners.

I just want to know that what is the effect of n-folds (in above example, we used 10-fold) on model. If we change n-fold, the performance of algorithm varies, how does it effect the performance?

```
kfold=model_selection.Kfold(n_splits=10, random_state=seed)
```



Jason Brownlee July 24, 2017 at 6:48 am #

REPLY ↩

The number of folds, and the specifics of the algorithm and data, will impact the stability of the estimated skill of the model on the problem.

Given a lot of data, often there is diminishing returns going beyond 10.

If in doubt, test the stability of the score (e.g. variance) by estimating model performance using a suite of different k values in k cross validation.



Nelson D'souza July 25, 2017 at 11:08 pm #

REPLY ↩

Hi! Jason,

Thanks for this amazing article/tutorial it is really very helpful.

I was working on a predictive model of my own

I seem to be occurring a problem nobody on the forum got 🤔 xD

I am sorry but could you help me out or point me in a direction ?

```
#####
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestClassifier
from sklearn import svm
from sklearn.svm import SVR

from sklearn import linear_model

import csv

from numpy import genfromtxt

import time
import datetime

from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

date = []
usage = []

date = genfromtxt('date.csv')
usage = genfromtxt('usage.csv')
test = genfromtxt('test.csv')

print (len(date))

print (len(usage))

dataframe = pd.DataFrame({
'Date': (date),
'Usage': (usage)
})
```

```

#drop NaN data's
dataframe = dataframe.dropna()
print (dataframe)

df = dataframe.drop(dataframe.index[[-1,-4]])

array = df.values

X = array[:,0:1]
Y = array[:,1]

validation_size = 0.20
seed = 7

X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)

seed = 7
scoring = 'accuracy'

# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()

```

```
#####
```

OutPut :

Date length : 366

Usage Length: 366

the data frame :

Date Usage

```

1 1.451587e+09 47139.0
2 1.451673e+09 85312.0
3 1.451759e+09 14301.0
4 1.451846e+09 20510.0
5 1.451932e+09 24225.0
6 1.452019e+09 30051.0
7 1.452105e+09 42228.0
8 1.452191e+09 27256.0
9 1.452278e+09 33746.0
10 1.452364e+09 30035.0
11 1.452451e+09 85844.0
12 1.452537e+09 28814.0
13 1.452623e+09 31082.0
14 1.452710e+09 21565.0
15 1.452796e+09 19095.0

```

```
16 1.452883e+09 15995.0
17 1.452969e+09 6578.0
18 1.453055e+09 96143.0
19 1.453142e+09 20503.0
20 1.453228e+09 31373.0
21 1.453315e+09 30776.0
22 1.453401e+09 39357.0
23 1.453487e+09 45955.0
24 1.453574e+09 21379.0
25 1.453660e+09 43682.0
26 1.453747e+09 51304.0
27 1.453833e+09 47333.0
28 1.453919e+09 33629.0
29 1.454006e+09 24185.0
30 1.454092e+09 47052.0
.. ... ..
336 1.480531e+09 74882.0
337 1.480617e+09 100712.0
338 1.480703e+09 45929.0
339 1.480790e+09 84837.0
340 1.480876e+09 85755.0
341 1.480963e+09 47184.0
342 1.481049e+09 62122.0
343 1.481135e+09 38140.0
344 1.481222e+09 46333.0
345 1.481308e+09 99399.0
346 1.481395e+09 101814.0
347 1.481481e+09 34078.0
348 1.481567e+09 45800.0
349 1.481654e+09 63657.0
350 1.481740e+09 33371.0
351 1.481827e+09 34921.0
352 1.481913e+09 33162.0
353 1.481999e+09 96179.0
354 1.482086e+09 27527.0
355 1.482172e+09 42291.0
356 1.482259e+09 112647.0
357 1.482345e+09 19299.0
358 1.482431e+09 52011.0
359 1.482518e+09 37571.0
360 1.482604e+09 78809.0
361 1.482691e+09 31469.0
362 1.482777e+09 69469.0
363 1.482863e+09 42879.0
364 1.482950e+09 31009.0
365 1.483036e+09 130637.0
```

[365 rows x 2 columns]

LR: 0.000000 (0.000000)

/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/discriminant_analysis.py:455:

UserWarning: The priors do not sum to 1. Renormalizing

UserWarning)

Traceback (most recent call last):

File "data_0.py", line 111, in

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/model_selection/_validation.py", line 140, in cross_val_score
for train, test in cv_iter)

File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 758, in __call__
while self.dispatch_one_batch(iterator):

File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 608, in dispatch_one_batch
self._dispatch(tasks)

File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 571, in _dispatch
job = self._backend.apply_async(batch, callback=cb)

```
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 109, in apply_async
result = ImmediateResult(func)
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/_parallel_backends.py", line 326, in __init__
self.results = batch()
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/model_selection/_validation.py", line 238, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/discriminant_analysis.py", line 468, in fit
self._solve_svd(X, y)
File "/Users/nelsondsouza/anaconda/lib/python2.7/site-packages/sklearn/discriminant_analysis.py", line 378, in _solve_svd
fac = 1. / (n_samples - n_classes)
ZeroDivisionError: float division by zero
```



Jason Brownlee July 26, 2017 at 7:55 am #

REPLY ↩

Sorry, I cannot debug your code. Consider posting to stackoverflow.



Nelson D'souza July 26, 2017 at 3:40 pm #

REPLY ↩

ok, Thanks 😊 Have a nice day!



Nelson D'souza July 26, 2017 at 6:49 pm #

REPLY ↩

I just thought I would let you know

my data set has 365 rows and only 2 columns is that a problem ?

Also I had a question, if you could lead me in a correct direction,

If my dataset has a column 'Dates' .datetime object how should I go about handling it ?

thanks in advance 😊



Jason Brownlee July 27, 2017 at 7:58 am #

REPLY ↩

Sounds like a time series forecasting problem. You should treat it differently.

Start here with time series forecasting:

<http://machinelearningmastery.com/start-here/#timeseries>



Soumya July 27, 2017 at 8:08 pm #

REPLY ↩

Awesome tutorial.. The program ran so smoothly without any errors. And it was easy to understand. Graphs looked fantastic.

Although I could not understand each and every functionality. Do you have any reference to understand the very basics of machine learning in Python?

Thanks for you help.



Jason Brownlee July 28, 2017 at 8:31 am #

REPLY ↩

Yes, start right here:

<http://machinelearningmastery.com/start-here/#python>



Razack July 29, 2017 at 3:46 pm #

REPLY ↩

Hi Jason,

Very nice tutorial. This helped me a lot.

Is there a way to append the train set with new data so that when ever I want I can add new data into the train model. What I could see creating new train sets.

Please help



Jason Brownlee July 30, 2017 at 7:39 am #

REPLY ↩

Not sure I follow.

Once you choose a model, you can fit a final model on all available data and start using it to make predictions on new data.

You may want to update your model in the future, in which case you can use the same process above with new data.

Does that help?



Dexter D'Silva August 2, 2017 at 11:34 pm #

REPLY ↩

Thank you Jason!!!

Having done the Coursera ML course by Andrew Ng I wasn't sure where to go next.

Your clear and well explained example showed me the way!!! Looking forward to reading your other material and spending many many more hours learning and having fun. (And my first foray into Python wasn't as daunting as I expected thanks to you).



Jason Brownlee August 3, 2017 at 6:51 am #

REPLY ↩

Thanks Dexter, well done on working through the tutorial!



Gerry August 3, 2017 at 5:51 am #

REPLY ↩

Hi Jason, I am using your tutorial for my own ML model and it's fantastic! I'm trying to predict make prediction on new data and am using

```
NB=GaussianNB()
```

```
new_prediction = predict.nb(new data)
```

```
print(new_prediction)
```

I am able to successfully get one prediction, how can I get the top 5 classifications for my new data? I have 15 possible classifications and I'd like the predict function to yield the top 5 instead of just the single prediction

Any help would be greatly appreciated, thank you so much!



Jason Brownlee August 3, 2017 at 6:57 am #

REPLY ↩

It sounds like your problem is a multi-class classification problem.

If so, you can predict probabilities and select the top 5 with the highest probability.

For example:

```
1 probabilities = model.predict_proba(X)
```



Gerry August 3, 2017 at 8:54 am #

REPLY ↩

Thanks, how can I match the probabilities to the class, or is there a way to have it return the class name?



Gerry August 3, 2017 at 9:08 am #

REPLY ↩

Here is the code:
ACN_prediction = NB.predict_proba([[0.80, 0.20, 0.70, 0.30, 0.99, 0.01, 0.98, 0.02, 0.95, 0.05, 0.95, 0.05, 1.00, 0]])
print (ACN_prediction)
And the result only displays:
[[0. 0. 0. ..., 0. 1. 0.]]
Is it just giving me the probabilities I have typed in?



Jason Brownlee August 4, 2017 at 6:44 am #

REPLY ↩

Each class is assigned an integer which is an index in the output array. This is done when you one hot encode the output variable.



Gerry August 3, 2017 at 9:30 am #

REPLY ↩

Using just the NB.predict([list of new data])
I would get the class 'Flower'
-Sorry for the long winded question, I have been stuck on this for hours, I appreciate your help



Jason Brownlee August 4, 2017 at 6:45 am #

REPLY ↩

If you just want one class label, then you do not need the probabilities and you can use predict() instead.



Gerry August 4, 2017 at 10:20 am #

REPLY ↩

If I want it to predict n best class labels I need to use predict_proba and manually match the n best probabilities to their class label correct? There is no other way to yield the top 5 class labels?



Jason Brownlee August 4, 2017 at 3:41 pm #

REPLY ↩

Yes. Correct.



Gerry August 5, 2017 at 6:10 am #

REPLY ↩

Thank you!



Jason Brownlee August 6, 2017 at 7:27 am #

REPLY ↩

I'm glad it helped.



Fernando D Mera August 10, 2017 at 1:54 am #

REPLY ↩

Hello, Jason,

I am using python3 on my mac, and I am also using Jupyter notebooks in order to complete the assignment on this webpage. Unfortunately, when I save the Iris dataset in my Desktop folder, and then run the command # shape
print(dataset.shape), the output is
(193, 5)

As you know, the output should be (150,5) and I am not sure why the dimensions of the dataset are wrong. Also, I tried to use the archive:
<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>, but the Jupyter output was the following

```
SSLError Traceback (most recent call last)
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/urllib/request.py in do_open(self, http_class, req, **http_conn_args)
1317 h.request(req.get_method(), req.selector, req.data, headers,
-> 1318 encode_chunked=req.has_header('Transfer-encoding'))
1319 except OSError as err: # timeout error

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in request(self, method, url, body, headers,
encode_chunked)
1238 """Send a complete request to the server."""
-> 1239 self._send_request(method, url, body, headers, encode_chunked)
1240

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _send_request(self, method, url, body, headers,
encode_chunked)
1284 body = _encode(body, 'body')
-> 1285 self.endheaders(body, encode_chunked=encode_chunked)
1286

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in endheaders(self, message_body, encode_chunked)
1233 raise CannotSendHeader()
-> 1234 self._send_output(message_body, encode_chunked=encode_chunked)
1235

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in _send_output(self, message_body, encode_chunked)
1025 del self._buffer[:]
-> 1026 self.send(msg)
1027

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in send(self, data)
963 if self.auto_open:
-> 964 self.connect()
965 else:

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py in connect(self)
1399 self.sock = self._context.wrap_socket(self.sock,
-> 1400 server_hostname=server_hostname)
1401 if not self._context.check_hostname and self._check_hostname:

How can I get the correct dimensions of the Iris dataset?
```



Jason Brownlee August 10, 2017 at 6:59 am #

REPLY ↩

Perhaps confirm that you downloaded the right dataset and have copied the code exactly.

Also, try running from the command line instead of the notebook. I find notebooks cause new and challenging faults.



Andrew Revoy August 14, 2017 at 7:39 am #

REPLY ↩

I've been eyeballing this tutorial for a while and finally jumped into it! I'd like to thank you for such a clear intro into machine learning! This has been the only tutorial I've found so far that actually has you evaluating the data / different models right off that bat.



Jason Brownlee August 15, 2017 at 6:26 am #

REPLY ↩

Thanks Andrew, and well done on working through it!



Abi Yusuf August 14, 2017 at 10:02 pm #

REPLY ↩

Hi Jason,

My sincere gratitude for this work you do to help us all out with ML. I have also been working away at this very wonderful field over the last 3 years now (PhD research – studying gaze patterns and trying to build predictive models of gaze patterns which represent some sort of behavior). In any case, I was reviewing the code you built here and I was just thinking that I don't tend to declare the test_size explicitly or the random_state either. I just put it directly into the algorithm

the random_state either – I just put it directly into the algorithm

so, your code goes:

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed) – totally spot on by the way,
```

My small addition/improvement – if you can call it that – would be to simply say

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size= 0.2, random_state= 7)
```

test_size keyword argument surely invokes the split method of the train_test_split module (I think) – meaning that the algorithm automatically assigns 80% to the training set and 20% to the test set

would you agree with this method? My python 3.x installation accepts this method just fine –

Also , I don't know if anyone else might have suggested this, but it is also worth pointing out that for cross_val (cv) – the fold size can be quite resource intensive and also there are underfitting/overfitting issues to be aware of, when doing cross validation –

Can you sense check these thoughts please?

Many Thanks.

Cheers



Jason Brownlee August 15, 2017 at 6:36 am #

REPLY ↩

Evaluating algorithms is an important topic.

Indeed the number of folds is important and we must ensure that each fold is sufficiently representative of the broader problem.

As for specifying the test size a different way, that's fine. Use whatever works best on your problem. The key is developing unbiased estimates of model skill on unseen data.



Sarbani August 15, 2017 at 5:08 am #

REPLY ↩

Thank you, Jason Brownlee, the post is very helpful. I was really lost in so many articles, blogs, open source tools. I was not able to understand how to start ML. Your post really helped me to start at least. I installed ANACONDA, ran the classification model successfully. Next Step – Understand the concept and apply on some real use cases.



Jason Brownlee August 15, 2017 at 6:44 am #

REPLY ↩

Well done Sarbani!



Ryan Stoddard August 15, 2017 at 3:39 pm #

REPLY ↩

Thanks for this extremely helpful example. I just have a question about your validation method as I was a little confused. It seems to me that you withhold 20% of the data for validation, then perform 10-fold cross-validation on only the 80% training data, then train a new model on entire 80% training data and test with 20% validation data. Is this correct, and if so is it common practice? It seems to me that the best way to get statistics about the best model is to simply use all of the data and perform 10-fold cross-validation. Why do you only perform cross-validation on 80% of the data, then evaluate a new model and only test it with a single validation set?



Jason Brownlee August 15, 2017 at 4:57 pm #

REPLY ↩

Great question Ryan!

We hold back a test set so that if we over fit the model via repeated cross validation (e.g. parameter tuning), we still have a final way of checking to see if we have fooled ourselves.

More here:

<http://machinelearningmastery.com/difference-test-validation-datasets/>

REPLY ↩



vishnu August 15, 2017 at 7:51 pm #

REPLY ↩

you above mention that scipy. it didn't available in pycharm (windows)..can u suggest another package for machine learning...?



Jason Brownlee August 16, 2017 at 6:33 am #

REPLY ↩

This tutorial will help you set up your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Adam Drake August 17, 2017 at 11:23 pm #

REPLY ↩

The link to download the "iris.dat" file appears to be broken!



Jason Brownlee August 18, 2017 at 6:20 am #

REPLY ↩

Here is the direct link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>



Ravindra Singh August 17, 2017 at 11:32 pm #

REPLY ↩

Thanks. Loved your result-first approach... Next I will use my own data set for a multi class problem. Hoping i would succeed !

A question

Given i will not have all the time to master writing new ML algorithms, I was wondering do i really need to ? I am an average developer from the past,(and new to Python but find it easy). I am thinking i should rather master how to prepare, present and interpret data – i understand domain very well – , and understand which algorithm (and libraries) to use for best results. I am guessing that, even to master applied ML, it will take many real projects !

I am keen in using ML in predicting data quality problems such as outliers that may need correction. any pointers ?



Jason Brownlee August 18, 2017 at 6:22 am #

REPLY ↩

Thanks Ravindra!

No, I recommend using a library, here's more on the topic:

<http://machinelearningmastery.com/dont-implement-machine-learning-algorithms/>

My best advice is to first collect a lot of data.



Brendan August 17, 2017 at 11:34 pm #

REPLY ↩

I am getting an error on the line starting with predictions?

```
# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

I am using Python 3, is there something else I need to install



Jason Brownlee August 18, 2017 at 6:24 am #

REPLY ↩

What error?

This tutorial will show you how to setup your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Ankith August 18, 2017 at 4:56 am #

REPLY ↩

Hey Jason!!!...Thanks for this!!!...Also I appreciate your helping out the people having doubts for, i guess an year!!! . I wish you good luck 😊



Jason Brownlee August 18, 2017 at 6:28 am #

REPLY ↩

Thanks Ankith, I'm glad the tutorial helped you.



fb August 18, 2017 at 9:54 am #

REPLY ↩

Thx a lot! Very helpfull



Jason Brownlee August 18, 2017 at 4:38 pm #

REPLY ↩

You're welcome.



beginner August 18, 2017 at 10:32 pm #

REPLY ↩

thank you this was really helpful >> too many indices for array
so I give him the data in 2 dimension instead of 1-D and use this >>> `numpy.loadtxt(dataset , delimiter=None , ndmin=2)` but he give me this error>>> could not convert string to float ,maybe because there are float and string in the iris file
what's the solution please I have to split them 😞
i'm really sorry for the bad english and thank you again <3



Jason Brownlee August 19, 2017 at 6:20 am #

REPLY ↩

Check your data file to makes sure it is a CSV file with no extra data.



beginner August 19, 2017 at 6:48 pm #

REPLY ↩

can you show me what do mean
my data file is the url you post it here, not an uploaded file
how can I do insure of this?(CSV file with no extra data)



Jason Brownlee August 20, 2017 at 6:05 am #

REPLY ↩

Use the filename or URL to load a file. It is that simple.



beginner August 18, 2017 at 10:44 pm #

REPLY ↩

Sorry I don't know where the rest of the previous comment disappeared>>so i a got a question
how could I separate the data such like this
`features = dataset[:,0:4]`
`classification = dataset[:,4]`
which is mean in other words when I write `print (dataset.shape)` I want him to give me :

(150,4) instead of (150,5) I told you that first I try to do this but he told me >> too many indices for array...continue reading at the beginning in the comment above



Xav August 19, 2017 at 3:03 am <#>

REPLY

I'd like to thank you for this concise but very helpful tutorial. I'm new to python and all the the code is clear apart the following part:

```
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

It's not clear to me how this 'for' cycle works. Specifically what is name and model?



Jason Brownlee August 19, 2017 at 6:23 am <#>

REPLY

It is evaluating the model using 10 fold cross validation. That means, 10 models are created and each is evaluated and the average score is calculated and stored in the list.

Does that help?



beginner August 19, 2017 at 7:19 am <#>

REPLY

did you mean to write this command?

```
dataset = pandas.read_csv(url, names = parameters)
```

I did like you do in this lecture and imported the data file from the link ,But still can not separate the data



Jason Brownlee August 20, 2017 at 6:03 am <#>

REPLY

What is the problem exactly?



Cole August 27, 2017 at 6:28 am <#>

REPLY

I think what he is trying to say is: he followed the tutorial as required, but once he got to the part where he had to load the iris dataset, he received a traceback from the line "dataset = pandas.read_csv(url, names = parameters)" in the python code provided. The traceback i received from this line was "NameError: name 'pandas' is not defined. Currently trying to fix, If i solve it before you get a chance to reply i will make sure to comment back on this tread what the problem was and how i fixed it.



Cole August 27, 2017 at 7:01 am <#>

REPLY

for section 2.2 to fix this error, imported panda along with the script. hopefully this did the trick. I do not understand why pandas needed to be imported again, but, i did it.

```
# Load dataset
import pandas
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)
print("its goin")
```



Jason Brownlee August 28, 2017 at 6:45 am #

REPLY ↩

Glad to hear it.



Jason Brownlee August 28, 2017 at 6:42 am #

REPLY ↩

It sounds like pandas is not installed.

This tutorial will help you install pandas and generally set-up your environment correctly:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Ernst August 20, 2017 at 8:29 am #

REPLY ↩

Wow. Great easy to use and understand example. It worked 100% for me. Thanks



Jason Brownlee August 21, 2017 at 6:04 am #

REPLY ↩

Thanks Ernst, I'm glad to hear that. Well done!



Dharik August 20, 2017 at 8:40 pm #

REPLY ↩

Hi Jason,

I found an error like this pls help me out.

```
# Compare Algorithms
```

```
... fig = plt.figure()
```

```
>>> fig.suptitle('Algorithm Comparison')
```



Jason Brownlee August 21, 2017 at 6:05 am #

REPLY ↩

Looks like a typo, change it to fig.subtitle()



Dharik August 22, 2017 at 5:01 pm #

REPLY ↩

But I copied it from your blog post.



Jason Brownlee August 23, 2017 at 6:42 am #

REPLY ↩

Oh, my mistake.



Dharik August 22, 2017 at 7:21 pm #

REPLY ↩

And I would like to create dataset, which is precisely focused on handwritten language recognition using RNN. Would you please share some of your ideas, thoughts and resources.



Jason Brownlee August 23, 2017 at 6:45 am #

REPLY ↩

Perhaps start here:

<https://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/>



Dharik August 24, 2017 at 3:50 pm #

Thank you Jason.



Jeremy August 25, 2017 at 1:16 am #

REPLY ↩

Awesome tutorial! Thanks Jason



Jason Brownlee August 25, 2017 at 6:44 am #

REPLY ↩

Thanks Jeremy.



Andrew August 25, 2017 at 2:50 am #

REPLY ↩

Hi Jason, in you post 5.1 Create a Validation Dataset. you wrote seed = 7.

What is seed and why did you choose #7?

Why not seed 10 or seed 5?

Andrew from Seattle



Jason Brownlee August 25, 2017 at 6:45 am #

REPLY ↩

Great question.

It does not matter what the value is as long as it is consistent.

See this post for a good explanation:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



ram August 30, 2017 at 7:48 pm #

REPLY ↩

Hi , this article is really nice.. I am executing statements..and those are also working fine..But still i am not getting what i am doing..I mean where is the logic? And what is this validation set means.What actually we are doing here? What is the intention?



Jason Brownlee August 31, 2017 at 6:17 am #

REPLY ↩

More on validation sets here:

<https://machinelearningmastery.com/difference-test-validation-datasets/>

More on the process of developing a predictive model end to end here:

<https://machinelearningmastery.com/start-here/#process>

Does that help?



KK SINGH September 1, 2017 at 4:08 am #

REPLY ↩

Hi jason,

Getting error in implementing

```
dataset.plot(kind='box', subplots=True, layout=(2, 2), sharex=False, sharey=False)
```

as:

```
super(FigureCanvasQT, self).__init__(figure=figure)
```

TypeError: 'figure' is an unknown keyword argument

Please help me.



Jason Brownlee September 1, 2017 at 6:51 am #

REPLY ↩

Might be an error in the way your environment is setup.

See this tutorial to setup your environment:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Ellie September 5, 2017 at 12:33 am #

REPLY ↩

Hi Jason!

When plotting the multivariate and univariate plots in Jupyter, I found them rather small. Is there a way to increase their size? I've tried using figsize, matplotlib.rcParams nothing seems to be working. Please help me out

Thanks!



Jason Brownlee September 7, 2017 at 12:36 pm #

REPLY ↩

Sorry, I don't use notebooks. I find them slow, hide errors and cause a lot of problems for beginners.



Kay September 6, 2017 at 11:11 pm #

REPLY ↩

Thank you, Jason.

Where in the model do you specify that you are predicting "class"? Did I miss that somewhere?



Jason Brownlee September 7, 2017 at 12:54 pm #

REPLY ↩

You can call `model.predict()`



Langue cedric September 8, 2017 at 2:12 am #

REPLY ↩

Very interesting.

That is my first tutorial on Machine learning.



Jason Brownlee September 9, 2017 at 11:46 am #

REPLY ↩

Thanks!



Sirish September 8, 2017 at 4:54 pm #

REPLY ↩

Dear Jason,

Firstly thank you very much for this wonderful blog.

i was trying this code on my project on a 8 lac rows data set

when tried

```
array = dataset.values
```

```
X = dataset.iloc[:, [0, 18]].values
```

```
y = dataset.iloc[:, 19].values
```

```
validation_size = 0.20
```

```
seed = 7
```

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

My Terminal gave me an error " positional indexers are out-of-bounds "

Summary of y data set is mentioned below

```
> print(dataset.shape)
```

```
> (787353, 18)
```

Could you pl help me in resolving this error



Jason Brownlee September 9, 2017 at 11:53 am #

REPLY ↩

Check your array slicing!



Garima Shrivastava September 8, 2017 at 11:21 pm #

REPLY ↩

Hi Jason

Grt work done by u.

I just completed this tutorial on python 2.7.1.but not able to predict the new class label using some new values



Jason Brownlee September 9, 2017 at 11:55 am #

REPLY ↩

Why not?



Albert September 11, 2017 at 3:22 am #

REPLY ↩

When doing the

```
# Load dataset
```

```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
```

```
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
```

```
dataset = pandas.read_csv(url, names=names)
```

section, terminal says

NameError: name 'pandas' is not defined

Is it that I don't have pandas installed correctly?



Jason Brownlee September 11, 2017 at 12:09 pm #

REPLY ↩

You need to install pandas.

See this tutorial:

<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Prashant September 12, 2017 at 2:34 am #

REPLY ↩

hi Jason....first of all thank for such a good tutorial.

my question is: while execution my python interpreter stuck at the following line:

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

and it neither produce any error nor correct output.

plz short it out...Thanks in advance.

I am using python 2.7.13



Jason Brownlee September 13, 2017 at 12:26 pm #

REPLY ↩



Perhaps wait a few minutes?



cesar September 13, 2017 at 5:14 pm #

REPLY ↩

Thank you so much Mr Joson, this tutorial is very helpful and professionally designed.
I also got this to ask, can we get the training time for each classifier produced?
The training vs testing error graph as well?
thank you again for the helping



Jason Brownlee September 15, 2017 at 12:00 pm #

REPLY ↩

I'm glad it helped.

Yes, you can develop these learning graphs, learn more here:
http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html



Trung Tiep September 13, 2017 at 6:37 pm #

REPLY ↩

Hi Jason,
seem this line of code doesn't work
`dataset.plot(kind = 'box', subplots = True, layout = (2,2), sharex = False, sharey = False)`
`plt.show()`
It doesn't show anything. Could you help me?
Thanks you and best regard



Jason Brownlee September 15, 2017 at 12:03 pm #

REPLY ↩

Are you able to confirm your environment is installed and working correctly:
<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>
Are you running the example as a Python script from the command line?



Dr. Pulak Mishra September 14, 2017 at 5:46 pm #

REPLY ↩

Traceback (most recent call last):
File "machinelearning1.py", line 63, in
`kfold = model_selection.Kfold(n_splits=10, random_state=seed)`
AttributeError: 'module' object has no attribute 'Kfold'

I have no idea about machine learning. just blindly following the tutorial example to just get an idea what is ML.
cn you tell me how am I supposed to correct this error.

I also wish you will be explaining all codes and functions in details step by step in future lessons



Jason Brownlee September 15, 2017 at 12:12 pm #

REPLY ↩

Looks like you might need to update your version of sklearn.

See this tutorial:
<http://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Chad September 15, 2017 at 2:47 am #

REPLY ↩

Hello Jason,

Thank you for your tutorial, it is amazing. Could you possibly do a follow up to this where you show how to package this, and use it? For instance I am not sure how to feed in new values, either manually or dynamically and then how could I store this data in a csv?



Jason Brownlee September 15, 2017 at 12:16 pm #

REPLY ↩

Great question.

I have some ideas about putting models into production here that might help as a start:

<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>



Silvio Abela September 16, 2017 at 1:29 am #

REPLY ↩

This is a superbly put tutorial for someone starting out in ML. Your step-by-step explanations allow people to actually understand and gain knowledge. Thank you so much for this and others that you have made.



Jason Brownlee September 16, 2017 at 8:42 am #

REPLY ↩

Thanks Silvio. Well done for working through it!



Niklas Wilke September 18, 2017 at 9:19 pm #

REPLY ↩

`dataset.hist()`

`plt.show()`

the 5&6 bar shows a different height on sepal-length ... did they change the dataset or anything? I'm not concerned, but just curious what could cause such a difference in display/result.

I imported everything properly, except the fact that I did not install theano because I'm planning to use TF. Can that have an issue on how it deals with data? Should I install it anyway?

<https://imgur.com/a/fC1TD>



Niklas Wilke September 18, 2017 at 10:20 pm #

REPLY ↩

Also I get different results when running my models... for me SVM is the best.

Could that be related to the visualization displaying something else before?

—Original—

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.981667 (0.025000)

—Original—

—Result—

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.975000 (0.038188)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

—Result—



Jason Brownlee September 19, 2017 at 7:44 am #

REPLY ↩

No, machine learning algorithms are stochastic

no, machine learning algorithms are stochastic.

Learn more here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Xinrui Li September 20, 2017 at 2:10 pm #

REPLY ↩

I also got SVM as the best model.

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.966667 (0.040825)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)



Jason Brownlee September 19, 2017 at 7:39 am #

REPLY ↩

That is odd, I don't have any ideas.



Niklas Wilke September 22, 2017 at 4:44 pm #

REPLY ↩

Could there be any changes to a newer version of the installed libraries ?

NumPy now working differently after they adjusted an algorythm or something like that ?

Maybe all who use the updated versions of all the included tools get this result ;/



Jason Brownlee September 23, 2017 at 5:36 am #

REPLY ↩

Machine learning algorithms are stochastic and generally give different results each time they are run:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Dan Harris September 23, 2017 at 4:27 pm #

REPLY ↩

Same here using python 3.6 (anaconda)

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.966667 (0.040825)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

Followed up with:

```
# Make predictions on validation dataset
```

```
svm = SVC()
```

```
svm.fit(X_train, Y_train)
```

```
predictions = svm.predict(X_validation)
```

```
print(accuracy_score(Y_validation, predictions))
```

```
print(confusion_matrix(Y_validation, predictions))
```

```
print(classification_report(Y_validation, predictions))
```

Resulting in:

```
0.933333333333
```

```
[[ 7 0 0]
```

```
[ 0 10 2]
```

```
[ 0 0 11]]
```

```
precision recall f1-score support
```

```
Iris-setosa 1.00 1.00 1.00 7
```

iris-setosa 1.00 1.00 1.00 1
Iris-versicolor 1.00 0.83 0.91 12
Iris-virginica 0.85 1.00 0.92 11
avg / total 0.94 0.93 0.93 30



Jason Brownlee September 24, 2017 at 5:14 am #

REPLY ↩

Nice work Dan!



Niklas Wilke September 27, 2017 at 6:38 pm #

REPLY ↩

you say they give out different results everytime , but it seems like everyone who is going through the tutorial right now is getting the "new" results.



Jason Brownlee September 28, 2017 at 5:23 am #

REPLY ↩

I tried to fix the random seed to make the example reproducible, but it is only reproducible within the set of libraries and their specific versions used. Even the platform can make a difference.



Jean Nunes September 26, 2017 at 6:06 am #

REPLY ↩

Hi, I'm new to machine learning. I started studying it for college purposes. Your tutorial really helped me and I was able to make it work with different datasets but now I wonder if there's a way, for example, to set the output (knn.__METHODNAME__('Iris-setosa')) and the method return generated data according to the parameter (in this case, sepal length and width and petal length and width). Thanks in advance!



Jason Brownlee September 26, 2017 at 2:58 pm #

REPLY ↩

You can make predictions for new observations by calling model.predict(X)

Does that answer your question?



delson September 28, 2017 at 4:05 pm #

REPLY ↩

hi sir ,can you help to make an artificial neural network on how i import my train data(weight ,biases)in python programming to classify its category in class 1 to 4 manually and input the sample as the program execute or run sir ,i have 5 neuron to test my Ai. thanks.



Jason Brownlee September 28, 2017 at 4:47 pm #

REPLY ↩

I have an example of coding a network from scratch here that you could use as a template:
<https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/>



Suresh Kmar September 29, 2017 at 12:28 am #

REPLY ↩

Great tutorial sir 😊

Im facing a problem in logistic regression with python +numpy +sklearn
How to convert all feature into float or numerical format for classification
Thanks



Jason Brownlee September 29, 2017 at 5:06 am #

REPLY ↩

You can use an integer encoding and a one hot encoding. I have many tutorials on the blog showing how to do this (use the search).



Keshav October 2, 2017 at 1:43 pm #

REPLY ↩

for me the result comes different:

LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.975000 (0.038188)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)

SVM is more accurate than KNN



vaibhav October 6, 2017 at 7:54 am #

REPLY ↩

same results. SVM is more accurate



Soumendra Kumar Dash October 3, 2017 at 1:56 am #

REPLY ↩

Hey

Nice guide. I did understand everything you have done but I had a small confusion regarding the seed variable being assigned to 7. I didn't understand its significance. Can you please tell me why we have considered the variable seed and why has it been assigned to 7 and not some other random number?



Jason Brownlee October 3, 2017 at 5:42 am #

REPLY ↩

It is to make the example reproducible.

You can learn more about the stochastic nature of machine learning algorithms here:
<https://machinelearningmastery.com/randomness-in-machine-learning/>



Abhijeet Singh October 3, 2017 at 5:40 pm #

REPLY ↩

In section 4.2 → Note the diagonal grouping of some pairs of attributes. This suggests a high correlation and a predictable relationship.

If u could explain how??



Jason Brownlee October 4, 2017 at 5:44 am #

REPLY ↩

Because the variables change together they appear as a line or diagonal line-grouping when plotted in 2D.



Nas October 3, 2017 at 11:15 pm #

REPLY ↩

File "ns.py", line 42

```
cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

^

IndentationError: unexpected indent

using my dataset I found this problem. How I can solve this type of problem please advice.



Jason Brownlee October 4, 2017 at 5:46 am #

REPLY ↩

Make sure you copy the code exactly.



Nas October 4, 2017 at 12:14 pm #

REPLY ↩

```
import pandas
from pandas.tools.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

dataset = pandas.read_csv("/home/nasrin/nslkdd/NSL_KDD-master/KDDTrain+.csv")

array = dataset.values
X = array[:,0:41]
Y = array[:,41]

validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, Y, test_size=validation_size, random_state=seed)

num_folds = 7
num_instances = len(X_train)
seed = 7
scoring = 'accuracy'

models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))

results = []
names = []
for name, model in models:
    kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
    cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring= Scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean()*100, cv_results.std()*100)
    print(msg)

.....

error is

Traceback (most recent call last):
File "ns.py", line 26, in
X_train, X_validation, Y_train, Y_validation = cross_validation.train_test_split(X, Y, test_size=validation_size, random_state=seed)
NameError: name 'cross_validation' is not defined
```



Jason Brownlee October 4, 2017 at 3:37 pm #

REPLY ↩

It looks like you might not have the most recent version of scikit-learn installed.



Yusuf October 5, 2017 at 10:52 am #

REPLY ↩

It's definitely the best site I've searched for machine learning. Thanks for everything!!

I wish you success in your business..



Jason Brownlee October 5, 2017 at 5:22 pm #

REPLY ↩

Thank you so much.



vaibhav October 6, 2017 at 7:52 am #

REPLY ↩

Hey, i am getting better results with the SVM algorithm, Why is it so? although we use the same data set.



Jason Brownlee October 6, 2017 at 11:03 am #

REPLY ↩

It is the stochastic nature of machine learning algorithms:

<https://machinelearningmastery.com/randomness-in-machine-learning/>

Also, there may have been changes to the library.



Amit October 6, 2017 at 5:03 pm #

REPLY ↩

Thanks Jason! its really beautiful to learn about ML . Thanks for your effort to make it effortless.



Jason Brownlee October 7, 2017 at 5:49 am #

REPLY ↩

Thanks Amit.



Davis October 8, 2017 at 12:26 am #

REPLY ↩

Thanks Jason its real great to do this project you open my eyes in the world of machine learning in python. Just have one question: how long does it take to learn algorithms in python?

and

is it advisable to learn python libraries for machine learning such as pandas, numpy, matplotlib and others before starting to learn different algorithms?



Jason Brownlee October 8, 2017 at 8:38 am #

REPLY ↩

You can make great progress in just a few weeks.

Yes, I recommend starting with Python, you can address a lot of practical problems. Get started here:

<https://machinelearningmastery.com/start-here/#python>



Kevin October 8, 2017 at 4:48 am #

REPLY ↩



Kevin October 6, 2017 at 4:46 am #

Does anyone offer Machine Learning tutoring? I need help and am having a hard time finding anyone willing to actually speak and talk through examples.



Jason Brownlee October 8, 2017 at 8:42 am #

REPLY ↩

I do my best on the blog 😊

Perhaps you can hire someone on upwork?



Praveen Kumar October 9, 2017 at 10:23 pm #

REPLY ↩

Hey Its really nice bu i have a question that for other kind of data sets is that procedure remains same..?



Jason Brownlee October 10, 2017 at 7:45 am #

REPLY ↩

It is a good start. Also see this more general procedure:

<https://machinelearningmastery.com/start-here/#process>



vinaya October 9, 2017 at 10:46 pm #

REPLY ↩

can you explain

```
X = array[:,0:4]
```

```
Y = array[:,4]
```



Jason Brownlee October 10, 2017 at 7:46 am #

REPLY ↩

We are selecting columns using array slicing in Python using ranges.

X is comprised of columns 0, 1, 2 and 3.

Y is comprised of column 4.



sukanya October 11, 2017 at 3:50 pm #

REPLY ↩

I am not clear with the seed value and its importance.can you expain this



Jason Brownlee October 11, 2017 at 4:41 pm #

REPLY ↩

It initializes the random number generator so that you get the same results as I do in the tutorial.

Generally, I recommend learning more about the stochastic nature of machine learning algorithms here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Ibrahim October 13, 2017 at 1:11 am #

REPLY ↩

Thanks Jason! its really beautiful to learn about ML using Python . Thanks for your effort to make it effortless. would you please recommend me unsupervised HMM using Python.

Thank you



Jason Brownlee October 13, 2017 at 5:49 am #

REPLY ↩



October 13, 2017 at 8:02 am #

Thanks. Sorry, I cannot help you with HMMs. I hope to cover the topic in the future.



Johnny October 13, 2017 at 8:02 am #

REPLY ↩

Why do you split the data into train and validation sets at the very beginning using "train_test_split"? I thought the K-Fold cross validation does that for us in this line:

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

I would assume we want to use the most data possible during model selection so why would we omit 20% of the data from this step?



Jason Brownlee October 13, 2017 at 2:53 pm #

REPLY ↩

We do this to double check the final model, learn more here:

<https://machinelearningmastery.com/difference-test-validation-datasets/>

Learn more about fitting a final model here:

<https://machinelearningmastery.com/train-final-machine-learning-model/>



Weizhi Song October 13, 2017 at 3:24 pm #

REPLY ↩

Hi Jason,

Thanks for your tutorial, it is really awesome! I want to use machine learning approach for biology problems. I have a question below and hope you could give me some suggestions. Thanks in advance.

I have eight DNA sequences which are labeled as either "TSS" or "NTSS". If I want to use your code here to predict whether a DNA sequence is TSS or not, do I need to transfer these sequences into numbers? If yes, do you have any suggestions of how to do that?

ATATATAG TSS
ACATTTAG TSS
ACATATAG TSS
ACTTATAG TSS
CCGTGTGG NTSS
CCGAGTGG NTSS
CCGTGCGG NTSS
CCGTCTGG NTSS

Thanks,
Weizhi



Jason Brownlee October 14, 2017 at 5:38 am #

REPLY ↩

Yes, you will need to encode each char or each block as an integer, and then perhaps as a binary vector.

See this post:

<https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>



Girmay October 13, 2017 at 10:51 pm #

REPLY ↩

This step by step tutorial is very interesting.

But I need yellow fever data set CSV file .. to predict yellow fever using machine learning.

Please anyone can help me...@ teklegimay@gmail.com



Jason Brownlee October 14, 2017 at 5:46 am #

REPLY ↩

Perhaps you can use google to find a suitable dataset?



Gaurav March 4, 2018 at 10:08 am #

REPLY ↩

go to ChEMBL dataset



Rash October 15, 2017 at 9:22 am #

REPLY ↩

Thanks for your help. This is awesome.

I have one issue : How can I rescale the axis ?

I have an error : ValueError: x and y must be the same size.

I have 3 features and 1 class for more than 245 000 data points.

please help.



Jason Brownlee October 16, 2017 at 5:40 am #

REPLY ↩

The error suggests that you must have the same number of input patterns as output labels.



Manish Sogi October 18, 2017 at 4:43 pm #

REPLY ↩

Hi Jason,

You might not be aware that your tutorial is arousing motivation to learn ML in engineers who are far away from this domain too. Thanks a ton !



Jason Brownlee October 19, 2017 at 5:33 am #

REPLY ↩

I'm glad to hear it!



Biswajith October 20, 2017 at 7:53 pm #

REPLY ↩

Hi Jason,

Nice and precise explanation. But can you please elaborate the problem definition here. Happy to see the step by step approach, still missing the actual problem or task we need to explore.

Below mentioned the basic stupid question.

What result are we expecting from this problem solution.

Biswa



Jason Brownlee October 21, 2017 at 5:33 am #

REPLY ↩

We are trying to predict the species given measurements of iris flowers.



shivaprasad October 24, 2017 at 4:46 am #

REPLY ↩

sir i am not getting what the classification report is ?, what is the meaning of precision, recall, f1 score and the support , what it actually tells us, what the table is for? , and what we understand with the help of the table



Jason Brownlee October 24, 2017 at 5:38 am #

REPLY ↩

Perhaps this article will help:

https://en.wikipedia.org/wiki/Precision_and_recall#Definition_.28classification_context.29



shivaprasad October 24, 2017 at 2:46 pm #

REPLY ↩

thank you sir



Micah October 25, 2017 at 3:58 am #

REPLY ↩

Great article. It's been a lot of help. I've been applying this to other free datasets to practice (e.g. the titanic dataset). One thing I haven't been able to figure out is how to show which columns are the most predictive. Do you know how to do that?

Thanks,
Micah



Jason Brownlee October 25, 2017 at 6:53 am #

REPLY ↩

Feature selection methods can give you an idea:

<http://machinelearningmastery.com/an-introduction-to-feature-selection/>



Daniel Bermudez October 26, 2017 at 8:48 am #

REPLY ↩

Hi Dr Jason,

I can't say thank you enough. This step by step tutorial is awesome. I'm so interested to try ML in a real project and this is a good way. I agree with you, academic is a little slow even though we can see more details.

Regards!!



Jason Brownlee October 26, 2017 at 4:15 pm #

REPLY ↩

I'm glad to hear it helped Daniel, well done for making it through the tutorial!



Aditya October 26, 2017 at 6:12 pm #

REPLY ↩

Sir,

I really appreciate your post and very thankful to you.
This post is very important for ML beginner like me.
I really loved the content and the way you make complex things simpler.

But I have one doubt, It would be very helpful to me if you help me building my understanding.

Question :

From the section "5.3 Build Models" line number 12

for name, model in models:

Please explain what is " name, model " here, its purpose and how it is working, (because I hadn't seen any FOR loop like this. I had learn python from YouTube videos and have very basic understanding)

P.S. I ran your code and its perfectly working fine.



Jason Brownlee October 27, 2017 at 5:18 am #

REPLY ↩

In that loop, a model is an item from the list, a "model" as the name suggests.

I recommend taking some more time to learn basic python loop structures:

<https://wiki.python.org/moin/ForLoop>



Aditya October 27, 2017 at 4:28 pm #

REPLY ↩

Thank you, you are awesome



Raj October 29, 2017 at 4:12 pm #

REPLY ↩

Hello Jason, I am curious about ai and ml. Tons of thanks for your hard work and commitment. I have done installation of Anaconda and checked all the libraries successfully. My ignorance of programming is compelling me to ask this ridiculous question. But i cant understand that where to upload dataset ? To be more clear i mean i dont understand even that where to write those url and given command to upload dataset ? on Jupiter notebook, or on conda prompt window ??? Please reply for kind of stupid question. Thanking you in anticipation.



Jason Brownlee October 30, 2017 at 5:36 am #

REPLY ↩

The function call `pandas.load_csv()` will load a CSV data file, either as a filename on your computer or a CSV file on a URL.

Does that help?



Kevin November 3, 2017 at 1:43 pm #

REPLY ↩

Thanks Jason! It's such a great article! However, i come across problems when applying your code here to my own dataset.

```
import sys
import scipy
import numpy
import pandas
import sklearn
```

```
from sklearn import model_selection
```

```
dataset = pandas.read_csv('D:\CMPE333\Project\Speed Dating Data_2.csv', header = 0)
```

```
array = dataset.values
```

```
X = array[:,0:12]
```

```
Y = array[:,12]
```

```
validation_size = 0.20
```

```
seed = 7
```

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

I got the error:

```
runfile('D:/CMPE333/Project/project.py', wdir='D:/CMPE333/Project')
```

Traceback (most recent call last):

File "", line 1, in

```
runfile('D:/CMPE333/Project/project.py', wdir='D:/CMPE333/Project')
```

File "C:\ProgramData\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 710, in runfile
execfile(filename, namespace)

File "C:\ProgramData\Anaconda3\lib\site-packages\spyder\utils\site\sitecustomize.py", line 101, in execfile
exec(compile(f.read(), filename, 'exec'), namespace)

File "D:/CMPE333/Project/project.py", line 33, in

```
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

AttributeError: module 'sklearn.model_selection' has no attribute 'train_test_split'

The dataset is stored as comma delimited csv file and has been loaded into a dataframe.

Can you tell me where is wrong? Thank you!!!



Jason Brownlee November 3, 2017 at 2:18 pm #

REPLY ↩

You might need to update your version of sklearn to 0.18 or higher.



Kevin November 4, 2017 at 6:35 am #

REPLY ↩

Thanks for replying!
My sklearn version is 0.18.1
It works well when i use your data.
Is there something wrong when i load the data?



Anil November 3, 2017 at 6:11 pm #

REPLY ↩

Hello Jason, Thank you. But one thing didn't clearly. Can you tell me in above example output what we predict? What we find? We are getting summarized the results as a final accuracy score, but about whos?



Jason Brownlee November 4, 2017 at 5:27 am #

REPLY ↩

We are predicting the iris flower species given measurements of flowers.



Meghal November 5, 2017 at 7:10 am #

REPLY ↩

Getting error in Class Distribution. If I give sum() instead of size() it works fine. Please suggest resolution.

```
=====
# class distribution
print(dataset.groupby('class').size())
=====

Output
Traceback (most recent call last):
File "C:\Python\ML\ImportLibs.py", line 30, in
print(dataset.groupby('class').size())
File "C:\Users\Meghal\AppData\Roaming\Python\Python35\site-packages\pandas\core\base.py", line 59, in __str__
return self.__unicode__()
File "C:\Users\Meghal\AppData\Roaming\Python\Python35\site-packages\pandas\core\series.py", line 1060, in __unicode__
width, height = get_terminal_size()
File "C:\Users\Meghal\AppData\Roaming\Python\Python35\site-packages\pandas\io\formats\terminal.py", line 33, in get_terminal_size
return shutil.get_terminal_size()
File "C:\Users\Meghal\AppData\Local\Programs\Python\Python35-32\lib\shutil.py", line 1071, in get_terminal_size
size = os.get_terminal_size(sys.__stdout__.fileno())
AttributeError: 'NoneType' object has no attribute 'fileno'
=====
```



Jason Brownlee November 6, 2017 at 4:44 am #

REPLY ↩

Perhaps double check you have the latest version of the libraries installed?
Confirm the data was loaded correctly?



Jeff Guo November 5, 2017 at 9:07 am #

REPLY ↩

Not sure why, but for me, SVM is giving me a higher accuracy in terms of precision, recall, and f1-score, but it ultimately has the same support score as KNN



Jason Brownlee November 6, 2017 at 4:47 am #

REPLY ↩

Might be the stochastic nature of ML algorithms:
<https://machinelearningmastery.com/randomness-in-machine-learning/>



xylo November 6, 2017 at 2:21 am #

REPLY ↩

1.can someone explain compare algorithm graph? 2.why knn is best algorithm 3. why & when use which algorithm?? thnx in advance



Jason Brownlee November 6, 2017 at 4:53 am #

REPLY ↩

Generally, we cannot know what algorithm will be “best” for a given problem. Our job is to use careful experiment to discover what works best for a given prediction problem.

See this post:

<http://machinelearningmastery.com/a-data-driven-approach-to-machine-learning/>



Georgios Koumakis November 7, 2017 at 4:48 am #

REPLY ↩

Jason, you are the best!!

Thanks for putting together all that material in a meaningful way, in a simple language and aesthetic environment.

There are not enough words to say how thankful I am.



Jason Brownlee November 7, 2017 at 9:53 am #

REPLY ↩

Thanks, I'm glad it helped Georgios.



Austin November 8, 2017 at 12:08 pm #

REPLY ↩

Hey Jason, fantastic tutorial. I have one questions though. Is there a way I could test the system by inputting a flower and the computer identifying it? Thank's a million!



Jason Brownlee November 9, 2017 at 9:52 am #

REPLY ↩

Yes, you could input the measurements of a new flower by calling `model.predict()`



Abhishek Jain November 9, 2017 at 1:36 am #

REPLY ↩

Hi Jason, Thanks a lot for the excellent step by step material to give a quick run-through of the methodology.

I am a tenured analytics practitioner and somehow found some time off to learn Python and was looking through the IRIS project itself. I had hypothesised that by adding more ratio variables to the dataset, we should get a better result on the prediction, Your excellent article gives me a ready code to test my hypothesis. I will share my results once I have them. 🙄



Jason Brownlee November 9, 2017 at 10:02 am #

REPLY ↩

Please do!



Abhishek Jain November 12, 2017 at 3:26 am #

REPLY ↩

Here are the k-Fold results: I used additional variables simply as all ratios of the original length variables respectively with no separate effort on dimensionality reduction.

LR: 0.950000 (0.040825)

LDA: 0.991667 (0.025000)
KNN: 0.958333 (0.055902)
CART: 0.950000 (0.066667)
NB: 0.966667 (0.055277)
SVM: 0.966667 (0.040825)

Drill down to the independent validation results for each technique:

Results for LR : 1.0

Results for LDA : 0.933333333333

Results for KNN : 1.0

Results for CART : 0.9

Results for NB : 0.966666666667

Results for SVM : 1.0

Although validation results are better across the board, I think LDA performs much better by this for K-fold method because other models may require a detailed variable selection or dimensionality reduction effort.

I would be glad to hear more from you on this. I am reachable on abhishek.zen@gmail.com.



Jason Brownlee November 12, 2017 at 9:06 am #

REPLY ↩

Great work, thanks for sharing!



narendra November 11, 2017 at 11:27 am #

REPLY ↩

Hi Jason,

Thank you for the great tutorial. once we run test and validate the model. How can we deploy the model. Also, how can we make the model predict on new data-set and still continuously learn from the new data.

Thank you,



Jason Brownlee November 12, 2017 at 9:00 am #

REPLY ↩

Great question.

This post has ideas on developing a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

This post has ideas on deploying a model:

<http://machinelearningmastery.com/deploy-machine-learning-model-to-production/>



chaitanya November 12, 2017 at 1:33 am #

REPLY ↩

Nice article to start with.

Although I really do not understand what each of model does?

So what should be the next step?



Jason Brownlee November 12, 2017 at 9:05 am #

REPLY ↩

You could learn more about how each model works:

<https://machinelearningmastery.com/start-here/#algorithms>



Anh November 13, 2017 at 9:15 pm #

REPLY ↩

Thanks a lot for your tutorial Jason. How should we apply the steps for Twitter data? Because the dataset is text, not number?



Jason Brownlee November 14, 2017 at 10:11 am #

REPLY ↩

Working with text is called natural language processing. You can get started with text here:
<https://machinelearningmastery.com/start-here/#nlp>



sanjay November 17, 2017 at 2:25 am #

REPLY ↩

"AxesSubplot" object has no attribute 'set_xticklables"



Jason Brownlee November 17, 2017 at 9:28 am #

REPLY ↩

Sorry to hear that, please confirm that you have setup your environment correctly:
<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>



Prateek Gupta November 17, 2017 at 11:20 pm #

REPLY ↩

Thanks Jason for this well explained post!
I am an aspiring data scientist and currently working on Walmart's sales forecasting dataset from kaggle.
If it is possible can you please also share a post about predicting the sales for this dataset?
It will be very helpful because I am not finding such a step by step tutorial in Python.



Jason Brownlee November 18, 2017 at 10:18 am #

REPLY ↩

Thanks for the suggestion.
Perhaps this process will help you work through the problem systematically:
<https://machinelearningmastery.com/start-here/#process>



ali November 20, 2017 at 3:58 pm #

REPLY ↩

Thanks for the amazing guide
can i know how to get the sensitivity and specificity and recall
you had a good Example Confusion Matrix in R with caret
but in the same page i could get the confusion for python but not the elements like
sensitivity and specificity and recall
thank again



Jason Brownlee November 22, 2017 at 10:37 am #

REPLY ↩

Perhaps this will help:
<http://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics>



Nicola November 22, 2017 at 6:09 am #

REPLY ↩

Thankyou very much for the great tutorial.
I analyzed every step but one thing it is not clear for me, and maybe it is the most important part of the tutorial 🤔
At the end of all our steps I would expect a function or something else to answer Python questions like these:
1. I have a flower with sepal-length=5, sepal width=3.5, petal-length=1.3 and petal-width=0.3, which class is it?
2. I have an Iris-setosa with sepal-length=5, sepal width=3.5, petal-length=1.3. What could be the petal-width?
Isn't this one of the the main objectives of the ML?



Nicola November 22, 2017 at 6:49 am #

REPLY ↩

OK, I answer by myself, for question one I could use

```
print(knn.predict([[5.0,3.5,1.3,0.3]]))
```

to get "['Iris-setosa']"

For question 2 I think that I need to rebuilt the whole model.



Jason Brownlee November 22, 2017 at 11:16 am #

REPLY ↩

Well done!



Jason Brownlee November 22, 2017 at 11:15 am #

REPLY ↩

Yes, you can train a final model on all data and use it to make a prediction.

Here's more about that:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

Here's how to save a model in Python:

<https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/>

You can predict on new data using:

```
1 X = ...  
2 yhat = model.predict(X)
```



Tash November 22, 2017 at 11:26 am #

REPLY ↩

This is a brilliant tutorial, thank you. I have a few questions – you split the data in to training and validation, but in this case would it not be classed as training and test?

Also, do you have any posts on tuning hyperparameters such as the learning rate in Logistic Regression? It was my understanding that a validation set would be used for something like this, while holding back the test set until the models been fine-tuned...but now I'm not sure if I'm confused!

Thanks so much.



Jason Brownlee November 23, 2017 at 10:23 am #

REPLY ↩

Yes, it would be training and test, here's more on the topic:

<https://machinelearningmastery.com/difference-test-validation-datasets/>



Túlio Campos November 24, 2017 at 11:56 am #

REPLY ↩

Why on

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

you use only the training part instead of the full set since it's a cross-validation?



Jason Brownlee November 24, 2017 at 3:05 pm #

REPLY ↩

In this case I wanted to hold back a test set to evaluate the final chosen model.



T lio Campos November 24, 2017 at 1:09 pm #

REPLY ↩

Also, in case I want to use X, Y by themselves. How could I arrange them in a ordered manner so I don't have totally random results because my classes aren't the right ones?

Thank you.



Jason Brownlee November 24, 2017 at 3:08 pm #

REPLY ↩

Sorry, I don't follow. Do do you have an example of what you mean?



T lio Campos December 5, 2017 at 3:29 am #

REPLY ↩

If you directly use

```
cv_results = model_selection.cross_val_score(model, X, Y, cv=kfold, scoring=scoring)
```

With kfold = 3 for example. You will get 3 different groups, each with one type of iris flower because sklearn doesn't shuffle it by its own and the dataset is arranged by flower-type.

You would have to use something like ShuffleSplit

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ShuffleSplit.html

Before doing so.



Jason Brownlee December 5, 2017 at 5:46 am #

REPLY ↩

Did you try this change, does it impact model skill as you suggest?



T lio Campos December 8, 2017 at 7:04 am #

Yes it does. In 3 fold I was getting under 70% accuracy. Shuffling makes it more evenly distributed (not 3 totally different groups). And I could get 90%_acc

Also, I figured that I could simply use the parameter "Shuffle=True" in .KFold



Jason Brownlee December 8, 2017 at 2:26 pm #

Nice!



Goldi November 25, 2017 at 12:30 pm #

REPLY ↩

Hi Jason,

Excellent way of explaining the basics of machine learning.

I assume that in almost all machine learning program if we are able to classify the data accurately then by applying algorithms we can understand much better about data .

classification is the key in supervised and clustering is the key in unsupervised learning is basics for a very good model.

Thanks a Lot.



Jason Brownlee November 26, 2017 at 7:30 am #

REPLY ↩

I'm glad you found it useful.



Meenakshi November 26, 2017 at 9:42 am #

REPLY ↩

Thanks for the tutorial, it is very helpful!



Jason Brownlee November 27, 2017 at 5:42 am #

REPLY ↩

You're welcome, I'm glad to hear that.



BENNAMA November 29, 2017 at 9:10 am #

REPLY ↩

I am working on windows 8.1

I am trying to apply the example by using python 2.7.14 anaconda

when arrived on section 4.1:

box and whisker plots

```
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
```

```
plt.show()
```

My cmd console shows an error "nameerror : plt name not defined"

To solve this problem i have added the line:

```
import matplotlib.pyplot as plt
```

it works

Thank's



Jason Brownlee November 30, 2017 at 8:04 am #

REPLY ↩

Glad to hear you fixed your issue.



Deepak Gautam December 2, 2017 at 5:23 am #

REPLY ↩

Hey! this is wonderful tutorial.

I goes through all the steps and it's great.

One thing I want to know that which is best model:-

* Linear Discriminant Analysis (LDA)

with 0.96

* K-Nearest Neighbors (KNN).

with 0.9



Jason Brownlee December 2, 2017 at 9:06 am #

REPLY ↩

It is up to the practitioner to choose the right model based on the complexity of the model and on mean and standard deviation of model skill results.



John Wolter December 4, 2017 at 10:04 am #

REPLY ↩

Here's a really nit-picky observation: You have two sections labeled 5.3.

Nit-picking aside, this is an excellent starter for ML in Python. I am currently taking the Coursera / Stanford University / Dr. Andrew Ng Machine Learning course and being able to see some of these algorithms that we have been learning about in action is very satisfying. Thank you!



Jason Brownlee December 4, 2017 at 4:57 pm #

REPLY ↩

Thanks John, fixed section numbering.



Ezra Axel December 5, 2017 at 4:50 pm #

REPLY ↩

How do you respond to all the comments?



Jason Brownlee December 6, 2017 at 8:59 am #

REPLY ↩

It takes time every single day!

But I created this blog to hang out with people just as obsessed with ML as me, so it's fun.



BukuBapi December 8, 2017 at 3:17 pm #

REPLY ↩

You Mentioned that

[We will use 10-fold cross validation to estimate accuracy.

This will split our dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits.]

In your code, I understand that you split it in 10 parts, but where is the 9:1 ratio mentioned. Unable to get that



Jason Brownlee December 9, 2017 at 5:36 am #

REPLY ↩

This is how cross-validation works, learn more here:

[https://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))



Nil December 11, 2017 at 12:39 am #

REPLY ↩

Hi Dr. Jason,

When evaluating we found that KNN presented the best accuracy, KNN: 0.983333 (0.033333). But when the validation set was used in KNN to have the idea of the accuracy, I see that the accuracy now is 0.9 so it decreased, while I was expecting the same accuracy. Can I consider this as over fitting? I can consider that KNN over fitted the train data? Is this difference of accuracy in the same model while training and validating acceptable?



Jason Brownlee December 11, 2017 at 5:26 am #

REPLY ↩

No, this is the stochastic variance of the algorithm. Learn more about this here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Nil December 11, 2017 at 9:37 pm #

REPLY ↩

Thank you.

I will learn more in the recommended site.

Best Regards.



bugtime December 11, 2017 at 5:21 am #

REPLY ↩

Jason,

AWESOME ARTICLE, THANK YOU!



Jason Brownlee December 11, 2017 at 5:34 am #

REPLY ↩

I'm glad it helped!



Gulshan Bhatia December 14, 2017 at 8:02 pm #

REPLY ↩

```
File "ml.py", line 73, in
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
File "/usr/local/lib/python2.7/dist-packages/sklearn/model_selection/_validation.py", line 342, in cross_val_score
pre_dispatch=pre_dispatch)
File "/usr/local/lib/python2.7/dist-packages/sklearn/model_selection/_validation.py", line 206, in cross_validate
for train, test in cv.split(X, y, groups))
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/parallel.py", line 779, in __call__
while self.dispatch_one_batch(iterator):
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/parallel.py", line 625, in dispatch_one_batch
self._dispatch(tasks)
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/parallel.py", line 588, in _dispatch
job = self._backend.apply_async(batch, callback=cb)
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/_parallel_backends.py", line 111, in apply_async
result = ImmediateResult(func)
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/_parallel_backends.py", line 332, in __init__
self.results = batch()
File "/usr/local/lib/python2.7/dist-packages/sklearn/externals/joblib/parallel.py", line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]
File "/usr/local/lib/python2.7/dist-packages/sklearn/model_selection/_validation.py", line 458, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)
File "/usr/local/lib/python2.7/dist-packages/sklearn/linear_model/logistic.py", line 1217, in fit
check_classification_targets(y)
File "/usr/local/lib/python2.7/dist-packages/sklearn/utils/multiclass.py", line 172, in check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'unknown'
```



Gulshan Bhatia December 14, 2017 at 8:08 pm #

REPLY ↩

urgent help required



Jason Brownlee December 15, 2017 at 5:31 am #

REPLY ↩

Confirm that you have copied all of the code and that your scipy/numpy/sklearn are all up to date.



Justin December 17, 2017 at 6:39 am #

REPLY ↩

Not sure if it's been mentioned, but this line: "pandas.read_csv(url, names=names)"
did not work for me until I replaced https with http after looking up docs for read_csv



Jason Brownlee December 17, 2017 at 8:55 am #

REPLY ↩

Thanks, Justin.



Nawaz December 19, 2017 at 7:59 pm #

REPLY ↩

hey Jason Brownlee,

Thanks for the tutorial

I got an error after I build five models

"urllib.error.URLError: "

Thanks



Jason Brownlee December 20, 2017 at 5:43 am #

REPLY ↩

Sorry to hear that. Perhaps ensure that your environment is up to date?



Zeinab December 20, 2017 at 4:42 pm #

REPLY ↩

Hello, Jason,

I am a beginner in python.

Unfortunately, when I load my dataset (it contains 4 features & 1 class "each with string datatype"), and then run the command
`cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring),`

I found the following error:

ValueError: could not convert string to float:



Jason Brownlee December 21, 2017 at 5:23 am #

REPLY ↩

Perhaps confirm that your data is all numerical?

Perhaps try converting it to float before using sklearn?



Steve H December 22, 2017 at 3:53 am #

REPLY ↩

Jason, great tutorial, this is extremely helpful! A couple of questions:

- 1) I realize that this is just an example, but in general, is this the process that you personally use when you are building production models?
- 2) What would the next steps be in terms of taking this to the next level? Would you choose the model that you think performs best, and then attempt to tune it to get even better results?



Jason Brownlee December 22, 2017 at 5:36 am #

REPLY ↩

Mostly, this is the process in more detail:

<https://machinelearningmastery.com/start-here/#process>



raymond doctor December 23, 2017 at 11:51 pm #

REPLY ↩

Hello,

The tutorial worked like a charm and I had no problem running it. However my need and that of a large number of linguists is different. As a linguist [and there are many like me throughout the world] we need to identify relationships within a source language or between a source and a target language.

At present I use an automata approach which states

`a->b` in environment `x`

This however implies that rules have to be manually written by hand and in the "brave new world" of big data this becomes a huge problem. I have searched and not located a simple tool which does this job using RNN. The existing tools are extremely complex and adapting them to suit a simple requirement of the type outlined above is practically impossible.

What I need is:

- a. A tool which installs itself deploying Python and all accompanying libraries.
- b. Asks for input of parallel data

- c. generates out rules in the back ground
- d. Provides an interface for testing by entering new data and seeing if the output works.
- e. It should work on Windows. A large number of such prediction tools are Linux based depriving both Windows and Mac users the facility to deploy them. My Windows10 is hopefully Linux Compatible but I have never tested the shell.
- f. Above all ease of use. A large number if not all Linguists are not very familiar with coding.

Do you know of any such tool ? And can such a tool be made available in Open Source. You would have the blessings of a large number of linguists who at present have to do the tedious task of generating out rules by hand and once again generating out new rules every time a sample not considered pops up.

I know the Wishlist above is quite voluminous.Hoping to get some good news

Best regards and thanks,

R. Doctor



Jason Brownlee December 24, 2017 at 4:54 am #

REPLY ↩

Sounds like an interesting problem. I'm not aware of a tool.

Do you have some more information on this problem, e.g. some links to papers or blog posts?



Prakash December 26, 2017 at 1:45 am #

REPLY ↩

Thanks for awesome tutorial....

I am facing issue in 4.1 section, while installing

```
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
```

I am getting this error.

Traceback (most recent call last):

File "", line 1, in

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 2677, in __call__

sort_columns=sort_columns, **kwds)

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 1902, in plot_frame

**kwds)

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 1729, in _plot

plot_obj.generate()

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 251, in generate

self._setup_subplots()

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_core.py", line 299, in _setup_subplots

layout_type=self._layout_type)

File "/usr/local/lib/python2.7/dist-packages/pandas/plotting/_tools.py", line 197, in _subplots

fig = plt.figure(**fig_kw)

File "/usr/local/lib/python2.7/dist-packages/matplotlib/pyplot.py", line 539, in figure

**kwargs)

File "/usr/local/lib/python2.7/dist-packages/matplotlib/backend_bases.py", line 171, in new_figure_manager

return cls.new_figure_manager_given_figure(num, fig)

File "/usr/local/lib/python2.7/dist-packages/matplotlib/backends/backend_tkagg.py", line 1049, in new_figure_manager_given_figure

window = Tk.Tk(className="matplotlib")

File "/usr/lib/python2.7/lib-tk/Tkinter.py", line 1818, in __init__

self.tk = _tkinter.create(screenName, baseName, className, interactive, wantobjects, useTk, sync, use)

_tkinter.TclError: no display name and no \$DISPLAY environment variable



Jason Brownlee December 26, 2017 at 5:18 am #

REPLY ↩

Sorry to hear that, looks like your Python installation may be broken.

Perhaps this tutorial will sort things out:

<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>

REPLY ↩



Rizwan Mian December 26, 2017 at 11:40 am #

Jason, I am learning so much from your work (thanks 🙏)

– my model scores are different to ones reported in the post (Section 5.4)? what could be the possible reasons?

('algorithm', 'accuracy', 'mean', 'std')

LR: 0.966667 (0.040825)

LDA: 0.975000 (0.038188)

KNN: 0.983333 (0.033333)

CART: 0.983333 (0.033333)

NB: 0.975000 (0.053359)

SVM: 0.991667 (0.025000)

– What do the bars represent in Algorithm Comparison in Section 5.4? Take LDA for example, the stated accuracy and standard deviation are 0.98 and 0.04. The bar in the chart finishes at about 0.94 and the whisker at about 0.92. Take knn for another example, the stated accuracy and standard deviation are 0.98 and 0.03. However, the bar finishes at 1 and the whisker at 0.92. How do I interpret the bars and whiskers? Is y-axis accuracy?

– how to read the confusion matrix without labels? My guess is row and column (missing) labels represent actual and predicted classes, respectively. However, I am unsure about the order of classes. is there a way to switch on the labels?

I collected and annotated the code in a python script (iris.py), and placed it on the github: <https://github.com/dr-riz/iris>



Jason Brownlee December 26, 2017 at 3:01 pm #

REPLY ↩

The differences may be related to the stochastic nature of the algorithms:

<https://machinelearningmastery.com/randomness-in-machine-learning/>

You can learn more about box and whisker plots here:

https://en.wikipedia.org/wiki/Box_plot

You can learn more about the confusion matrix here:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>

Great annotations, please reference the URL of this blog post and the name of the blog as source.



Rizwan Mian December 27, 2017 at 7:16 am #

REPLY ↩

Thanks for your reply and reminder. Credits and Source,URLs are now noted in README. 🙏

Re LDA example: the stated accuracy and standard deviation are 0.98 and 0.04. Yes, the box plot renders metrics such as minimum, first quartile, median, third quartile, and maximum but *not* necessarily mean. Hence, we don't see mean and std in the box plot in Section 5.4.

I reproduce this with a simple example.

```
lda_model = LinearDiscriminantAnalysis()
```

```
lda_results = model_selection.cross_val_score(lda_model, X_train, Y_train, cv=10, scoring='accuracy')
```

np.size(lda_results) => 10 elements, 1 for each fold. Shouldn't it for every test sample?separate investigation.

```
lda_results.max() # => 1
```

```
numpy.median(lda_results) # > 1
```

```
numpy.percentile(lda_results, 75) # => 1 — 3rd quartile
```

```
numpy.percentile(lda_results, 25) # => 0.9423 — 1st quartile: 0.94230769230769229
```

```
lda_results.min() # => 0.9091 — this is value whisker we see
```

```
lda_results.mean() # => 0.9749 — DONT expect to see in the plot
```

```
lda_results.std() # => 0.03849 — DONT expect to see in the plot
```

```
fig = plt.figure()
```

```
ax = fig.add_subplot(111)
```

```
plt.boxplot(lda_results)
```

```
ax.set_xticklabels(['LDA'])
```

```
plt.show()
```

As expected, we don't see mean and std in the box plot.



Jason Brownlee December 28, 2017 at 5:18 am #

REPLY ↩

Thanks.

Cross validation is creating 10 models and evaluating each on 10 different and unique samples of your dataset.



Daniel December 28, 2017 at 9:12 am #

REPLY ↩

Nice. Took me a little longer than 10 mins, but works as advertised. (I did everything under python3, no big difference I think.)

What would be really cool here would be a "what is going on here" section at the end. But it's real nice to have something that actually runs, and be able to poke about with it a bit.

Thanks Jason. Good stuff.



Jason Brownlee December 28, 2017 at 2:10 pm #

REPLY ↩

Well done. Nice suggestion, thanks.



MG5 December 29, 2017 at 3:26 am #

REPLY ↩

Hello Jason, I wanted to ask you if the seed dataset can be treated like iris, using your tutorial I arrived at 97% accuracy, do you think it can still improve? The dataset site is: <https://archive.ics.uci.edu/ml/datasets/seeds>.



Jason Brownlee December 29, 2017 at 5:25 am #

REPLY ↩

Perhaps, though that is an impressive result.



Sammy Lee December 29, 2017 at 12:38 pm #

REPLY ↩

So how would we obtain individual new predictions using our own input data after going through this exercise?



Jason Brownlee December 29, 2017 at 2:37 pm #

REPLY ↩

Train a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>

Then call:

```
1 X = ...
2 yhat = model.predict(X)
```



Gage Russell December 29, 2017 at 3:35 pm #

REPLY ↩

I am getting the syntax error pasted below at the start of the for loop to evaluate each model. I have made sure that I am copying and pasting it directly, and tried a few of my own fixes. Any help as to why this is occurring would be great! Thanks in advance!

for name, model in models:

File "", line 1

for name, model in models:

^

SyntaxError: unexpected EOF while parsing



Jason Brownlee December 30, 2017 at 5:17 am #

REPLY ↩



Ensure that you copy all of the code with the same formatting. White space has meaning in Python.



Joe January 1, 2018 at 10:00 am #

REPLY ↩

I put the requirements for this tutorial in a Dockerfile if anyone is interested: <https://github.com/UnitasBrooks/docker-machine-learning-python>



Jason Brownlee January 2, 2018 at 5:31 am #

REPLY ↩

Thanks Joe.



Rizwan Mian January 1, 2018 at 2:21 pm #

REPLY ↩

The algorithms are instantiated with their default parameters. Is this a standard practise for spot checking algorithms?



Jason Brownlee January 2, 2018 at 5:33 am #

REPLY ↩

You can specify some standard or common configurations as part of the checking.



abidh January 1, 2018 at 6:32 pm #

REPLY ↩

I tried the above tutorial. But i got accuracies differ from the given above for the same dataset. why? also the boxplot for the same is changing each time



Jason Brownlee January 2, 2018 at 5:34 am #

REPLY ↩

Yes, this is a feature not a bug, learn more here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Ben Hart January 6, 2018 at 5:01 pm #

REPLY ↩

Hi Jason,

I think I downloaded the same dataset as you have here but the sepal-length data seems to have changed a bit. Not to worry though as you can easily follow the exact same steps except you just have to make predictions using SVC ()

LR: 0.966667 (0.040825)
LDA: 0.975000 (0.038188)
KNN: 0.983333 (0.033333)
CART: 0.966667 (0.040825)
NB: 0.975000 (0.053359)
SVM: 0.991667 (0.025000)

0.933333333333

[[7 0 0]

[0 10 2]

[0 0 11]]

precision recall f1-score support

Iris-setosa 1.00 1.00 1.00 7

Iris-versicolor 1.00 0.83 0.91 12

Iris-virginica 0.85 1.00 0.92 11

avg / total 0.94 0.93 0.93 30

It does give a better result which is nice.

Also I was wondering if you explain the confusion matrix anywhere on your website, I find it somewhat confusing 🙄



Jason Brownlee January 7, 2018 at 5:04 am #

REPLY ↩

Yes, here is more on the confusion matrix:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



dj January 6, 2018 at 7:01 pm #

REPLY ↩

What we predicted in the output with help of iris dataset



Jason Brownlee January 7, 2018 at 5:04 am #

REPLY ↩

The model predicts the species based on flower measurements.



Praveen Chakravarthy January 7, 2018 at 10:53 pm #

REPLY ↩

Hi Jason, watched your videos and you are awesome, can you tell me how to train our own image data database and split into train and test sets, labels...thank you for listening to me...



Jason Brownlee January 8, 2018 at 5:43 am #

REPLY ↩

I don't have any videos.



prageeth January 8, 2018 at 10:57 pm #

REPLY ↩

Thank you so much..



Jason Brownlee January 9, 2018 at 5:30 am #

REPLY ↩

You're welcome.



Jackson January 10, 2018 at 3:34 am #

REPLY ↩

Hi Jason,

Thanks for this great tutorial. It really helps.

Everything works fine except:

a. In Section 4.1 – Histogram – the distribution in Sepal Length is quite different from yours. May be that's due to the random nature of Machine Learning ?

b. In section 5.4 – Box and whisker plot: the plots for LR, LDA and CART are similar but for KNN, SVM; I could only get a "+" sign at around 0.92 (no box and no whisker shown). For NB, I could only get 1 "+" sign at 0.92 and 1 "+" sign at around "0.83".

Grateful if you could advise. Thanks.

I am using :

window 10, python 3.5.2 – Anaconda custom (64 bit)

scipy: 1.0.0

numpy: 1.13.3

matplotlib: 1.5.3

python 2.7.11

pandas: 0.18.1
statsmodels: 0.6.1
sklearn: 0.19.1

theano: 0.9.0.dev-unknown-git
Using TensorFlow backend.
keras: 2.1.2



Jason Brownlee January 10, 2018 at 5:30 am #

REPLY ↩

Well done!



Jackson January 11, 2018 at 2:45 am #

REPLY ↩

Thanks, but something goes “wrong”. Grateful if you could advise.

In section 5.4 – Box and whisker plot: the plots for LR , LDA and CART are similar to that shown in your web page

but for KNN, SVM; I could only get a “+” sign at around 0.92 (no box and no whisker shown). For NB, I could only get 1 “+” sign at 0.92 and 1 “+” sign at around “0.83”.



Jason Brownlee January 11, 2018 at 5:53 am #

REPLY ↩

Interesting.



NAVALUTI SHIVAKUMAR January 13, 2018 at 6:02 am #

REPLY ↩

thank you so much for valuable blog.

I'm new to Python and ML. your blog is helped me a lot in learning.

in this I've not understand how data will train (X_train , Y_train and)

thanks



Jason Brownlee January 13, 2018 at 7:49 am #

REPLY ↩

Thanks.



Chandi January 15, 2018 at 9:29 pm #

REPLY ↩

Hello Jason,

This is amazing tutorial and it's really helps me to understand well!!!.. Please I want to know, do you have this type of tutorials for “pyspark” ? Can you suggest me any links, books, pdf or any tutorials? Thank you



Jason Brownlee January 16, 2018 at 7:33 am #

REPLY ↩

Not at this stage, sorry.



Nilotpal January 16, 2018 at 2:19 pm #

REPLY ↩

It has a dependency with pillow library, but it is not mentioned, or did I miss something?



Jason Brownlee January 17, 2018 at 9:55 am #

REPLY ↩

Does it?

Perhaps this is contingent on how you setup your environment?



EDUARDO DURAN January 23, 2018 at 4:00 pm #

REPLY ↩

Dear ,

Maybe you have the .py file of the tutorial? could you send it to me please



Jason Brownlee January 24, 2018 at 9:51 am #

REPLY ↩

It is a part of this book:

<https://machinelearningmastery.com/machine-learning-with-python/>



Jude January 26, 2018 at 12:08 am #

REPLY ↩

Thank you, Jason Brownlee. I did run the entire scripts. It worked simply well on my MacBookPro. You are the best!



Jason Brownlee January 26, 2018 at 5:43 am #

REPLY ↩

I'm glad to hear it, well done Jude!



Sunil January 27, 2018 at 4:55 am #

REPLY ↩

Hi Jason,

Very nice tutorial.

I am getting error while running models. It is complaining about reshaping the data.

Following is the stacktrace

Traceback (most recent call last):

File "C:\eclipse_workspace\MachineLearning\Iris_Project\src\IrisLoadData.py", line 86, in trainData()

File "C:\eclipse_workspace\MachineLearning\Iris_Project\src\IrisLoadData.py", line 30, in trainData
run_algorithms(X_train, Y_train, seed, scoring)

File "C:\eclipse_workspace\MachineLearning\Iris_Project\src\IrisLoadData.py", line 79, in run_algorithms
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)

File "C:\Python27\lib\site-packages\sklearn\model_selection_validation.py", line 342, in cross_val_score
pre_dispatch=pre_dispatch)

File "C:\Python27\lib\site-packages\sklearn\model_selection_validation.py", line 206, in cross_validate
for train, test in cv.split(X, y, groups))

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 779, in __call__
while self.dispatch_one_batch(iterator):

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 625, in dispatch_one_batch
self._dispatch(tasks)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 588, in _dispatch
job = self._backend.apply_async(batch, callback=cb)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib_parallel_backends.py", line 111, in apply_async
result = ImmediateResult(func)

File "C:\Python27\lib\site-packages\sklearn\externals\joblib_parallel_backends.py", line 332, in __init__
self.results = batch()

File "C:\Python27\lib\site-packages\sklearn\externals\joblib\parallel.py", line 131, in __call__
return [func(*args, **kwargs) for func, args, kwargs in self.items]

File "C:\Python27\lib\site-packages\sklearn\model_selection_validation.py", line 458, in _fit_and_score
estimator.fit(X_train, y_train, **fit_params)

```
estimator.fit(X_train, y_train, **fit_params)
```

File "C:\Python27\lib\site-packages\sklearn\linear_model\logistic.py", line 1216, in fit
order="C")

File "C:\Python27\lib\site-packages\sklearn\utils\validation.py", line 573, in check_X_y
ensure_min_features, warn_on_dtype, estimator)

File "C:\Python27\lib\site-packages\sklearn\utils\validation.py", line 441, in check_array
"if it contains a single sample.".format(array))

ValueError: Expected 2D array, got 1D array instead:

```
array=[2.8 3. 3. 3.3 3.1 2.2 2.7 3.2 3.1 3.4 3.8 3. 3.3 2.4 2. 2.8 3.4 2.9
```

```
3.5 3.1 2.9 2.6 2.7 4.4 3.2 3.4 4. 2.6 2.5 3. 3. 3.2 2.9 3. 3. 3.8
```

```
3.2 3.2 3. 2.6 2.4 3.1 4.2 3. 3.2 3.5 3.8 2.8 2.9 3.7 2.5 3.4 2.8 3.
```

```
3.2 3.7 3.3 2.8 2.5 2.8 2.3 3.4 3.9 2.8 3. 3.7 2.7 3.2 3.4 2.8 2.3 3.1
```

```
3.1 3.6 3. 2.9 2.8 2.8 3.1 2.9 3. 2.7 3. 2.3 2.8 3.4 3.3 2.5 3.8 3.8
```

```
3.4 2.8 3. 3.5 3. 3. 2.2 3.4 3.2 3.2 2.5 2.5 3.3 2.7 2.6 2.9 2.7 3. ].
```

Reshape your data either using `array.reshape(-1, 1)` if your data has a single feature or `array.reshape(1, -1)` if it contains a single sample.

Could you please take a look and help me out?



Jason Brownlee January 27, 2018 at 5:59 am #

REPLY ↩

Perhaps double check your loaded data meets your expectations?



Sunil January 28, 2018 at 5:16 am #

REPLY ↩

Hi Jason,

Yeah I made some mistake while loading the data. I corrected it.

I have some questions.

What is confusion matrix and support in final result? Can you please tell about these things? For logistic regression/ classification algorithms, we need to calculate weights and we need to provide learning rate for cost function and we need to minimize it right? Is it taken care in python libraries?

Thank you,
Sunil



Jason Brownlee January 28, 2018 at 8:27 am #

REPLY ↩

See this post on the confusion matrix:

<https://machinelearningmastery.com/confusion-matrix-machine-learning/>



Pythor January 27, 2018 at 2:16 pm #

REPLY ↩

This was fun for my first Machine learning project. I was stuck on making pygames since I learned Python



Jason Brownlee January 28, 2018 at 8:21 am #

REPLY ↩

Well done!



Gopal Venugopal January 28, 2018 at 9:58 am #

REPLY ↩

Hello,

I have a technical problem please! I have downloaded Anaconda 3.6 for windows in my desktop. However, I am unable to see Terminal window or Anaconda Prompt although I have the anaconda navigator installed. Is there something wrong?

Thank you very much for your advise,

Gopal.



Jason Brownlee January 29, 2018 at 8:14 am #

REPLY ↩

Perhaps this post will help:

<https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>



Jenny January 29, 2018 at 5:33 pm #

REPLY ↩

I just want to say thank you this is very helpful!



Jason Brownlee January 30, 2018 at 9:47 am #

REPLY ↩

You're welcome, glad to hear that.



kotrappa SIRBI January 30, 2018 at 12:39 pm #

REPLY ↩

Very nice Machine Learning getting started like HelloWorld, Thanks



Jason Brownlee January 31, 2018 at 9:36 am #

REPLY ↩

I'm glad it helped.



Blessy January 30, 2018 at 3:57 pm #

REPLY ↩

i get this error after the line

```
" cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring) "
```

Traceback (most recent call last):

File "", line 1, in

File "C:\Users\HP\AppData\Local\Programs\Python\Python36-32\lib\site-packages\sklearn\model_selection_validation.py", line 335, in cross_val_score

scorer = check_scoring(estimator, scoring=scoring)

File "C:\Users\HP\AppData\Local\Programs\Python\Python36-32\lib\site-packages\sklearn\metrics\scorer.py", line 274, in check_scoring
"fit" method, %r was passed" % estimator)

TypeError: estimator should be an estimator implementing 'fit' method, [(('LR', LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,

intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,

penalty='l2', random_state=None, solver='liblinear', tol=0.0001,

verbose=0, warm_start=False)), ('LDA', LinearDiscriminantAnalysis(n_components=None, priors=None, shrinkage=None,

solver='svd', store_covariance=False, tol=0.0001)), ('KNN', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',

metric_params=None, n_jobs=1, n_neighbors=5, p=2,

weights='uniform')), ('CART', DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,

max_features=None, max_leaf_nodes=None,

min_impurity_decrease=0.0, min_impurity_split=None,

min_samples_leaf=1, min_samples_split=2,

min_weight_fraction_leaf=0.0, presort=False, random_state=None,

splitter='best')), ('NB', GaussianNB(priors=None)), ('SVM', SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,

decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf',

max_iter=-1, probability=False, random_state=None, shrinking=True,

tol=0.001, verbose=False))] was passed



Jason Brownlee January 31, 2018 at 9:37 am #

REPLY ↩

Sorry to hear that, I have not seen this error. Perhaps try updating your libraries?



Rahul January 31, 2018 at 5:52 pm #

REPLY ↩

Sorry, If its a very basic question. I am a newbie in Machine Learning. Was trying to understand the explanation.

I have a question at below code block, where we are splitting the dataset into input (X) and output(Y). What is the use of the output set ? What is its significance ?

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```



Jason Brownlee February 1, 2018 at 7:16 am #

REPLY ↩

The output is the thing being predicted.

This post might help you understand how algorithms work:

<http://machinelearningmastery.com/how-machine-learning-algorithms-work/>



Rahul February 1, 2018 at 6:19 pm #

REPLY ↩

Jason, one more more clarification needed on the "output values" . In many articles , I have seen that ML works only on numeric values (even its of different type we need to convert it to numeric). Doesn't it apply to the "output values" we are using ? Don't we need to convert them to numeric ?



Jason Brownlee February 2, 2018 at 8:09 am #

REPLY ↩

Generally, yes we do.



Bipin Singh January 31, 2018 at 8:43 pm #

REPLY ↩

Great article for beginners. Thanks you very much. Jason do you have any more articles for more in depth knowledge?



Jason Brownlee February 1, 2018 at 7:19 am #

REPLY ↩

Yes, start here:

<https://machinelearningmastery.com/start-here/>



Ityav Luke February 1, 2018 at 1:20 pm #

REPLY ↩

Sir,

Through your article i have successfully installed python 2.7 anaconda and every stage i got it successful. Now as i tried to delve into this tutorial i am problems.

I first run a check on versions of libraries as you said and the result is okay:

```
Python: 2.7.14 [Anaconda custom (64-bit)] (default, Oct 15 2017, 03:34:40) [MSC
v.1500 64 bit (AMD64)]
scipy: 0.19.1
numpy: 1.13.3
matplotlib: 2.1.0
```

pandas: 0.20.3
sklearn: 0.19.1

The next step which is to import libraries and i did by copy and pasting into a script file running with this command: python script.py and not error shown.

Where i had problem is to load the dataset csv from ML repo.

As i execute the command to load dataset from a script file

i have the following error

Traceback (most recent call last):

File "script.py", line 4, in

```
dataset = pandas.read_csv(url, names=names)
```

NameError: name 'pandas' is not defined

Please what is the issue here?

thanks



Jason Brownlee February 2, 2018 at 8:04 am #

REPLY ↩

Perhaps you have two versions of Python installed accidentally?



Rahul February 1, 2018 at 6:11 pm #

REPLY ↩

Got it now.

If i am correct, the initially supplied output values gives the model an inference that for some given set of inputs, this would be the output ?
And finally, based on this my model will be trained and then work on the entirely new inputs provided to the system ?



Jason Brownlee February 2, 2018 at 8:08 am #

REPLY ↩

Sorry, I don't follow.



Bipin Singh February 1, 2018 at 7:45 pm #

REPLY ↩

Just a minor suggestion which i encountered, pandas.tools.plotting is deprecated,
use pandas.plotting instead.

Thanks 🙏



Jason Brownlee February 2, 2018 at 8:16 am #

REPLY ↩

Thanks, fixed.



chanid February 1, 2018 at 8:44 pm #

REPLY ↩

Hello Jason,

I'm always fan of your tutorials. Please, have done any tutorials like this for explaining every algorithm in depth including mathematics behind them, how and what exactly happening in side the algorithm.

Thank you



Jason Brownlee February 2, 2018 at 8:18 am #

REPLY ↩

I have two books that explain how algorithms work:

<https://machinelearningmastery.com/products>



Martine February 2, 2018 at 8:25 pm #

REPLY ↩

Hello,

I get this error:

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/multiclass.py in check_classification_targets(y)
170 if y_type not in ['binary', 'multiclass', 'multiclass-multioutput',
171 'multilabel-indicator', 'multilabel-sequences']:
-> 172 raise ValueError("Unknown label type: %r" % y_type)
173
174
```

ValueError: Unknown label type: 'continuous'

I am using my own dataset. What is wrong here?



Jason Brownlee February 3, 2018 at 8:35 am #

REPLY ↩

Perhaps your dataset is the problem?



Hugues Laliberte February 4, 2018 at 7:12 am #

REPLY ↩

Hi Jason,

i'm also using my own dataset, and i get the same error as Martine above:

File "/Users/Hugues/anaconda3/lib/python3.6/site-packages/sklearn/utils/multiclass.py", line 172, in check_classification_targets
raise ValueError("Unknown label type: %r" % y_type)
ValueError: Unknown label type: 'continuous'

I can check my dataset, but what should we be looking for ? I have used that dataset with the LSTM model without any error messages.

thanks



Hugues Laliberte February 4, 2018 at 7:16 am #

REPLY ↩

The multiclass.py code that is giving the error is:

```
if y_type not in ['binary', 'multiclass', 'multiclass-multioutput',
'multilabel-indicator', 'multilabel-sequences']:
raise ValueError("Unknown label type: %r" % y_type)
```

line 172 is the last line

looks like 'continuous' is not expected. Where is 'continuous' coming from ?



Hugues Laliberte February 4, 2018 at 7:19 am #

REPLY ↩

my last column is binary, 0 or 1



Hugues Laliberte February 4, 2018 at 7:32 am #

REPLY ↩

googling this error code i find the following solution:

"You are passing floats to a classifier which expects categorical values as the target vector."

I thought my last column is categorical because it contains only 1 and 0, but i guess i0'm wrong. Is there a way out ?



Huques Laliberte Februarv 4. 2018 at 7:37 am #

REPLY ↩

i changed my last column from 0 and 1 to 'zero' and 'one'
now the error message changes to:
ValueError: Unknown label type: 'unknown'
I'm getting closer....



Jason Brownlee February 5, 2018 at 7:40 am #

REPLY ↩

Sorry, I have not seen this error before. Perhaps try posting to stackoverflow?



Hugues February 6, 2018 at 1:20 am #

REPLY ↩

i found the problem now. This part of your code above has to be changed according to the number of columns of our data set:

```
1 # Split-out validation dataset
2 array = dataset.values
3 X = array[:,0:4]
4 Y = array[:,4]
5 validation_size = 0.20
6 seed = 7
7 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

So the 4 in X and Y needs to be changed. This seems obvious now but i'm new to Python and this is a rather dense language.

thanks a lot, the best output for my data set is KNN with 85%. I will now try to improve on this by cleaning my data.



Jason Brownlee February 6, 2018 at 9:19 am #

REPLY ↩

Why does it need to be changed?



jcridge February 7, 2018 at 4:17 am #

REPLY ↩

Please change Section 2.1 out of date reference

CURRENT TEXT

from pandas.plotting import scatter_matrix

TO REVISED TEXT

from pandas.tools.plotting import scatter_matrix

as per comments already submitted

thanks



Jason Brownlee February 7, 2018 at 9:28 am #

REPLY ↩

The "pandas.tools.plotting" is outdated.

The latest version of Pandas uses "pandas.plotting".

Consider updating your version of Pandas to v0.22.0 or higher.

Learn more here:

<https://pandas.pydata.org/pandas-docs/stable/visualization.html#scatter-matrix-plot>



Phil February 7, 2018 at 5:01 am #

REPLY ↩

Hi Jason

Apologies if this has already been asked.

What would be the next step, therefore, if I wanted to apply this prediction to new data? I.e. if we got a new data set with just the measurements, how do we program the use of the predictions we've found to estimate the species?

P.s. great blog, really useful!



Phil February 7, 2018 at 5:06 am #

REPLY ↩

ah don't worry, you can just apply `knn.predict()` to a new array of the sizes right? That's easy



Jason Brownlee February 7, 2018 at 9:31 am #

REPLY ↩

Correct.

Also see this post on creating a final model:

<https://machinelearningmastery.com/train-final-machine-learning-model/>



jcridge February 8, 2018 at 1:50 am #

REPLY ↩

RE: is the validation dataset nugatory given the k-fold validation process

Whilst the idea of separating out a "final independent test data set (30 samples)" away from the k-fold cross validation process seems nice, is it not actually wasting the opportunity to develop and compare the N model types using the larger and therefore more useful data set within the k-fold process ?

In short, the k-fold process seems to already be doing everything that the hold-out sample is purporting to do.

Out another way, surely the hold out data is no more independent than the i(th) hold out data partitioned within i(th) k-fold execution ?



Jason Brownlee February 8, 2018 at 8:31 am #

REPLY ↩

There are many approaches at estimating out of sample model skill. I recommend finding an approach that is robust for your specific problem.



Pallavee February 9, 2018 at 6:28 pm #

REPLY ↩

Hello Jason,

This post is a great starting point – I am new to coding (with only basics at hand), python with lot of interest in ML. The post has got me started with it... I was able to run most of the tutorial successfully with few experiments by changing the graphs, seed values, k-folds etc. Few questions though –

1. In one of the answers you have explained how kfold works on February 17, 2017 –
Now in the for loop, where you define kfold for a model at hand, that split is done only once right? I mean e.g. for LR, being first model to evaluate, we split the data of 120 in 10 folds with 12 items in each. Then as explained in the above post – The model is trained on the first 9 folds and evaluated on the records in the 10th. When we go for next set of 9, we are NOT resplitting the 120 items in new 10 sets right?
2. Also, when you say model is trained on first 9 folds – It means that we are looking at the relationships of the 4 numeric values and the class (out of 3 – Iris-setosa, Iris-versicolor, Iris-virginica) which they belong to, right?
3. When the dataset is split between X and Y values (Y being the output/ result of relationships between 4 values in X), where in the code are we actually mentioning this? I mean how/ where does the algorithm gets to know that X are the independent variables and Y is the dependent variable in which we want to classify our data?

Thanks a lot!

Pallavee



Jason Brownlee February 10, 2018 at 8:55 am #

REPLY ↩

No, the same split into folds is reused with a new model fit and evaluated on different sets each time, systematically.

Yes, a fit model really means a learned mapping from inputs to outputs.

res, a fit model really means a learned mapping from inputs to outputs.
<http://machinelearningmastery.com/how-machine-learning-algorithms-work/>

We specify the inputs and outputs to the model as separate parameters in sklearn.



Raghavendra February 9, 2018 at 9:03 pm #

REPLY ↩

Hi Jason,

I am getting below errors.

Statement: from pandas.plotting import scatter_matrix
throws error as "No module named plotting"

Statement: from sklearn import model_selection
throws error as "cannot import name model_selection"

Regards
Raghavendra



Jason Brownlee February 10, 2018 at 8:55 am #

REPLY ↩

You will need to update your version of pandas and sklearn to the latest versions.



Bipin February 9, 2018 at 9:34 pm #

REPLY ↩

Hi Jason on my dataset I used kfold but couldn't find any significant difference. Can you explain why this may happen. Also, does using kfold cross_validation lead to overfitting?

P.S:

with cross_validation without cross_validation
LogisticRegression 0.816 0.816
LinearDiscriminantAnalysis 0.806 0.806
KNeighborsClassifier 0.79 0.79
DecisionTreeClassifier 0.810 0.816
GaussianNB 0.803 0.803
SVC 0.833 0.833
LinearSVC 0.806 0.806
SGDClassifier 0.7525 0.620
RandomForestClassifier 0.833 0.803



Jason Brownlee February 10, 2018 at 8:56 am #

REPLY ↩

Both do the same job of performing k-fold cross validation.

You can overfit when evaluating models with cross validation, although it is less likely on average than using other evaluation methods.



Akheel February 10, 2018 at 6:36 pm #

REPLY ↩

Excellent tutorial Jason, and thanks very much for it.

One noob question here though –

Where do 'dataset' and 'plt' get associated in the code above? I ask this coz I don't see any code where we are associating 'dataset' and 'plt'; and yet when we call 'plt.show()', the plot that gets drawn has data from the 'dataset'.



Jason Brownlee February 11, 2018 at 7:53 am #

REPLY ↩

The dataset is loaded:

```
1 dataset = pandas.read_csv(url, names=names)
```

plt is the pyplot library

```
1 import matplotlib.pyplot as plt
```

A search on the page (control-f) would have helped you discover this for yourself.



Akheel February 13, 2018 at 12:51 am #

REPLY ↩

Thanks Jason, but that i know.

Let me try to make my question clearer –

From the examples I studied to understand pyplot, the recurring idea is

1. set the range to be plotted along the x-axis [let's says that's e]
2. provide the corresponding values to be plotted along the y-axis [let's say that's f]
3. Steps 1 and 2 are accomplished by the call – 'plt.plot(e, f)'
4. After the call to 'plot', the call to 'show' is made which will display the plot

ex:

```
e = np.arange(0.0, 2.0, 0.01)
f = 1 + np.sin(2*np.pi*t)
plt.plot(e, f)
plt.show()
```

As you can see, the call to 'plot' provides the values to 'plt' and the call to 'show' will cause the plotting and display of the same from 'plt'.

However, in your example, I don't see any line which is equivalent to the 'plot' call.

So my question is – When and where does 'plt' get the values from 'dataset' that it uses to draw the plot?

I hope it's clearer now.



Jason Brownlee February 13, 2018 at 8:04 am #

REPLY ↩

Here, I use pandas to make the calls to matplotlib via the pandas DataFrame (called dataset), then call plt.show().



Mr D February 11, 2018 at 7:58 am #

REPLY ↩

I installed Anaconda according to your instructions (<https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/>) but as I go to run python and check the versions of libraries I get this:

```
... import numpy
```

Traceback (most recent call last):

File "", line 2, in

ImportError: No module named numpy

How can I get passed this.



Jason Brownlee February 12, 2018 at 8:25 am #

REPLY ↩

It looks like numpy is not installed or you are trying to run code in a different version of Python from anaconda.



Najmath February 13, 2018 at 3:45 pm #

REPLY ↩

Hello Jason,

I have a project in which it should predict the disease by specifying the symptoms.How can I implement this and can you please help me with the attributes of symptoms and all.



Jason Brownlee February 14, 2018 at 8:13 am #

REPLY ↩

I recommend this process:

<https://machinelearningmastery.com/start-here/#process>



pradnya February 13, 2018 at 4:33 pm #

REPLY ↩

Thank you very much jason... for the great tutorial.
its really great aratical...its help so much to our project..thanks...



Jason Brownlee February 14, 2018 at 8:14 am #

REPLY ↩

I'm glad it helped.



Cor Colijn February 16, 2018 at 10:10 am #

REPLY ↩

Hi Jason,

Well I got the example running but only after I deleted "scoring=scoring" in code below:

for name, model in models:

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
```

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

```
results.append(cv_results)
```

```
names.append(name)
```

```
msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
```

With "scoring=scoring" I received error message something like "scoring not defined".

Then when I added "scoring=scoring" back I did not received the error and the program runs fine.

What could this be?

Anyhow, great tutorial.

Regards,
Cor



Jason Brownlee February 16, 2018 at 2:57 pm #

REPLY ↩

Glad to hear you overcame your issue.

you might have missed a snippet from earlier in the example where "scoring" was assigned.



Akshata February 16, 2018 at 4:49 pm #

REPLY ↩

Hi Jason,

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

After typing that line in my command prompt, it shows this error:

Traceback (most recent call last):

File "", line 1, in

NameError: name 'model' is not defined

I tried copy pasting that line directly offthe tutorial, I still faced the same error. What should I do??



Jason Brownlee February 17, 2018 at 8:40 am #

REPLY ↩

I think you may have missed some lines of code from the tutorial.



Cor Colijn February 16, 2018 at 11:52 pm #

REPLY ↩

I did get this exact error also. Then when I removed "scoring=scoring", thinking 'well, maybe the compiler or whatever is smart enough to deal with this', the code worked as expected. Then when I reinserted "scoring=scoring", I did not get the error message and the code continued to run as expected.



feedsack February 17, 2018 at 2:49 am #

REPLY ↩

When I run this code

```
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

i get this error

TypeError: cannot perform reduce with flexible type

and i get a blank graph where x_axis and y_axis both are labelled from 0.0-1.0 at every 0.2 interval.

How do I fix it?



Jason Brownlee February 17, 2018 at 8:49 am #

REPLY ↩

Sorry, I have not seen this fault, perhaps post to stackoverflow?



mufassal February 19, 2018 at 3:37 am #

REPLY ↩

what algorithm should i use for weather prediction



Jason Brownlee February 19, 2018 at 9:09 am #

REPLY ↩

As far as I know, modern weather forecasting uses physical models, not machine learning methods.

That being said, if you do want to explore ML methods for weather forecasting, I would recommend this process:

<https://machinelearningmastery.com/start-here/#process>



John Bagiliko February 21, 2018 at 9:52 pm #

REPLY ↩

from pandas.plotting import scatter_matrix

That did not work until I used

from pandas import scatter_matrix

Maybe this can help someone also.



Jason Brownlee February 22, 2018 at 11:17 am #

REPLY ↩

Interesting, perhaps you need to update your version of Pandas?

Here is the API for "pandas.plotting.scatter_matrix":

<https://pandas.pydata.org/pandas-docs/stable/visualization.html#scatter-matrix-plot>



Bob Fujita February 22, 2018 at 11:56 am #

REPLY ↩

Just started your tutorial. Looks like the best introduction to machine learning. I'm getting the following error while trying to load the iris dataset. Would appreciate your assistance in correcting my problem. Thanks.

```
===== RESTART: /Users/TinkersHome/Documents/load_data.py =====
>>> dataset = pandas.read_csv(url, names=names)
Traceback (most recent call last):
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/urllib/request.py", line 1318, in do_open
    encode_chunked=req.has_header('Transfer-encoding'))
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1239, in request
    self._send_request(method, url, body, headers, encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1285, in _send_request
    self.endheaders(body, encode_chunked=encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1234, in endheaders
    self._send_output(message_body, encode_chunked=encode_chunked)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1026, in _send_output
    self.send(msg)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 964, in send
    self.connect()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/http/client.py", line 1400, in connect
    server_hostname=server_hostname)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 407, in wrap_socket
    _context=self, _session=session)
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 814, in __init__
    self.do_handshake()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 1068, in do_handshake
    self._sslobj.do_handshake()
File "/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/ssl.py", line 689, in do_handshake
    self._sslobj.do_handshake()
ssl.SSLError: [SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed (_ssl.c:777)
```



Jason Brownlee February 23, 2018 at 11:51 am #

REPLY ↩

Sorry, I have not seen this error. Perhaps try searching/posting on stackoverflow for the error message?



Angela February 22, 2018 at 8:56 pm #

REPLY ↩

Hello experts,

When practise 5.Algorithm, I encountered this error message. Also checked all the installed tools & packages, which are all up-to-date. Kindly please help me to fix it, thanks very much.

```
>>> # Spot Check Algorithms
... models = []
>>> models.append(('LR', LogisticRegression()))
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
```



```
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'msg' is not defined
```



Jason Brownlee February 23, 2018 at 11:56 am #

REPLY ↩

Ensure that you copy all of the code for the example and that your indenting matches the example in the tutorial.



Angela February 23, 2018 at 8:38 pm #

REPLY ↩

I will retry. Thank you very much Jason. Cheers!



Jason Brownlee February 24, 2018 at 9:11 am #

REPLY ↩

Hang in there!



Alan February 22, 2018 at 11:32 pm #

REPLY ↩

Hi Jason,

Great tutorial, thanks!

I got an unique error that no one had posted here – special...

The error is at this line:

```
cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
```

And it says: ValueError: This solver needs samples of at least 2 classes in the data, but the data contains only one class: 0.0

But my X_train.shape shows (52480L, 25L) and my y_train.shape is (52480L,).

Any ideas please?

Thanks,

Alan



Jason Brownlee February 23, 2018 at 11:58 am #

REPLY ↩

Hi Alan, it means that your data does not have enough examples in each class.

The dataset may be highly imbalanced.

If so, this post might give you some ideas:

<https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>

REPLY ↩



Bob Fujita February 23, 2018 at 5:31 am #

REPLY ↩

Added the following lines to my load dataset file & now all is well:

```
import ssl
ssl._create_default_https_context = ssl._create_unverified_context
```



Jason Brownlee February 23, 2018 at 12:03 pm #

REPLY ↩

Nice one!



isaias February 26, 2018 at 1:14 am #

REPLY ↩

Hello, Mr Jason!

I'm learning ML and PLN and i have a lot of doubts:

you can recommend some article, blog (and so on) to learn more about this? I have to implement a model switching different classifiers for predict/discriminate a class. The model is described below:

- I have a set S of words;
- Each word W of S is a class for prediction;

Two different of vector of features are used:

- 1 – The first is a vector which use PMI score between W and n-gram occurring before W and PMI between W and n-gram placed after W. Then, the vector length is twice length of S (set of words);
- 2 – Other is a vector of 500 most words (vocabulary) occurring in a context (variable size) surrounding all words of S. If the word (feature) exists in a sentence for training, the vector puts '1' or '0', otherwise. Frequency of word on document (context/sentence) don't matter here.

I know that i have to vectorize features and create a array of counts, but i can't understand even a little about what way i've to follow after that steps (roughly explained).

Basically, above informations are the most important.

Finally, i wanna use the different classifiers in a "plugable" way. Its possible?

Thanks in advance.



Jason Brownlee February 26, 2018 at 6:05 am #

REPLY ↩

My best advice for getting started with NLP is here:

<https://machinelearningmastery.com/start-here/#nlp>



Phillip C. February 26, 2018 at 11:43 pm #

REPLY ↩

Great tutorial!

In my case, I am POSTing the IRIS data to a Flask web service, but I don't see how to get that data into a pandas dataframe using any of the "read_csv" or other methods available. I tried to use io.String(csv_variable), then using read_csv on that, but it still doesn't work.

Suggestions?

Thanks,



Jason Brownlee February 27, 2018 at 6:32 am #

REPLY ↩

Perhaps try posting the question to stackoverflow?



Griffin February 27, 2018 at 2:14 am #

REPLY ↩

Hi Jason!

First of all, great introduction to cross validation! Your tutorial is comprehensive and I appreciate that you went through everything step-by-step as much as possible.

Just a question regarding section 5.3 Build Models. This was taken from your code directly:

```
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
```

As I have looked at other websites on cross validation as well, I am confused on the X and y inputs. Should it be X_train and Y_train or X and Y (original target and data)? Because I looked at sklearn documentation (http://scikit-learn.org/stable/modules/cross_validation.html#cross-validation), it seems that the original target and data were used instead, and they did not perform a train_test_split to obtain X_train and Y_train.

Please clarify. Thank you!



Jason Brownlee February 27, 2018 at 6:36 am #

REPLY ↩

The goal in this part is to evaluate the skill of the model. The data would be the training data, a sample of data from your domain.

Perhaps this post would clear things up for you:

<https://machinelearningmastery.com/difference-test-validation-datasets/>



Ron February 28, 2018 at 1:19 pm #

REPLY ↩

What is the main objective of this project?



Jason Brownlee March 1, 2018 at 6:06 am #

REPLY ↩

To teach you something.

The model will learn the relationship between flower measurements and iris flower species. Once fit, it can be used to predict the flower species for new flower measurements.



anushri February 28, 2018 at 7:51 pm #

REPLY ↩

I believe there are many more pleasurable opportunities ahead for individuals that looked at your site.



Jason Brownlee March 1, 2018 at 6:12 am #

REPLY ↩

Thanks.



Attharuddin March 6, 2018 at 6:14 am #

REPLY ↩

for name, model in models:

```
kfold = model_selection.KFold(n_splits=10, random_state=seed)
cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
results.append(cv_results)
names.append(name)
msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
print(msg)
```

I could not run this code, please help me out



Jason Brownlee March 6, 2018 at 6:21 am #

REPLY ↩

Why not? What was the problem?



Christian Post March 6, 2018 at 10:43 pm #

REPLY ↩

Great example to see what you can and can't do with your data.

I ran this with my own sample and well, did not get over 70% accuracy so it looks like my data is just not good 🙄

I just had to do some small adjustment since this line is hard-coded:

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

I had to change it because my dataset has only 3 independent variables:

```
# Split-out validation dataset
array = dataset.values
n = dataset.shape[1]-1
X = array[:,0:n]
Y = array[:,n]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

I think this should work regardless of the number of attributes in any given dataset(?)



Jason Brownlee March 7, 2018 at 6:14 am #

REPLY ↩

Nice.



mahima kapoor March 7, 2018 at 1:43 am #

REPLY ↩

i need to build a taxi passenger seeking system using machine learning, i am a beginner. how should i go about it? please suggest some relevant source codes for reference



Jason Brownlee March 7, 2018 at 6:15 am #

REPLY ↩

Perhaps this process will help:

<https://machinelearningmastery.com/start-here/#process>



Pauli Isoaho March 10, 2018 at 8:49 am #

REPLY ↩

Excelnt guide, thank you

What enviroment you need to plot?



Jason Brownlee March 11, 2018 at 6:16 am #

REPLY ↩

Thanks.

What do you mean by environment?



Nick F March 10, 2018 at 8:43 pm #

REPLY ↩

Thanks for the tutorial. When I run the code, the Support Vector Machine got the best score (precision 0.94), while the knn got precision 0.90, as in your example. I am using Python 3. Is the different result caused by the global warming? 🤔



Jason Brownlee March 11, 2018 at 6:24 am #

REPLY ↩

Nice work.

A difference in results is caused by the stochastic nature of the algorithms:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Frank984 March 10, 2018 at 9:55 pm #

REPLY ↩

I have Python: 2.7.10 (default, May 23 2015, 09:40:32) and the following versions of the libraries:

scipy: 0.15.1
numpy: 1.9.2
matplotlib: 1.4.3
pandas: 0.16.2
sklearn: 0.18.1

I have modified your example considering the following structure for the dataset:

```
Age Weight Height Metbio RH Tair Trad PMV TSV gender
0 61 61.4 175 2.14 31.98 21.35 20.58 -0.38 0 male
1 39 81.0 178 2.19 46.88 24.25 24.09 0.30 1 male
[...]
```

All works fine, except for the following part:

I have created a validation dataset considering:

```
# Split-out validation dataset
array = dataset.values
X = array[:,0:8]
#the line above is interpreted as "all rows for columns 0 through 8"
Y = array[:,9]
#the line above is interpreted as "all rows for column 9"
validation_size = 0.20
# 20% as a validation dataset
seed = 7
#what does this parameter means?
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
```

Now when I try to built and evaluate the 6 models with this code:

```
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

It appears this message:

```
>>> # Spot Check Algorithms
... models = []
>>> models.append(('LR', LogisticRegression()))
```

```

>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
...     kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model' is not defined
>>> results.append(cv_results)
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'msg' is not defined
>>>

```

Could you explain how can I solve?



Frank984 March 10, 2018 at 10:24 pm #

REPLY ↩

I have tried also anaconda prompt and the following versions:

```

Python 3.6.1 |Anaconda 4.4.0 (64-bit)| (default, May 11 2017, 13:25:24)
scipy: 0.19.0
numpy: 1.12.1
matplotlib: 2.0.2
pandas: 0.20.1
sklearn: 0.18.1

```

Same error when I try to build and evaluate the six models considering the script of paragraph 5.3



Jason Brownlee March 11, 2018 at 6:26 am #

REPLY ↩

Versions look ok. Ensure you have all proceeding code for each example.



Jason Brownlee March 11, 2018 at 6:26 am #

REPLY ↩

Looks like a copy-paste error.

Ensure you copy all of the code and maintain the same indenting.



Frank984 March 12, 2018 at 5:51 am #

REPLY ↩

Solved considering this post:

<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/#comment-431754>



Kevin March 13, 2018 at 10:47 am #

REPLY ↩

Hi Jason,

Your Instruction were great. I am new to coding and I would like to know if you have codes for fantasy sports. Will the process above work with fantasy sports.



Jason Brownlee March 13, 2018 at 3:05 pm #

REPLY ↩

Not at this stage. I have worked on sports datasets using rating systems and had great success:

https://en.wikipedia.org/wiki/Elo_rating_system



Qasem March 13, 2018 at 9:57 pm #

REPLY ↩

how long will it take to run the program? i follow all instruction, and there is no errors, but still running and only get the first graph, and the dataset description? is it take to long to complete run ? note i use windows 7



Jason Brownlee March 14, 2018 at 6:20 am #

REPLY ↩

Seconds. No more than minutes.



Qasem March 14, 2018 at 12:08 pm #

REPLY ↩

so what do you think is the problem?



Qasem March 14, 2018 at 12:27 pm #

REPLY ↩

I have done like this and its just work till # histograms, there problem the pycharm 3 does not show any error.

```
# Load libraries
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC

# Load dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)

# shape
print(dataset.shape)

# head
print(dataset.head(20))

# descriptions
print(dataset.describe())
```

```

print(dataset.describe())
# class distribution
print(dataset.groupby('class').size())
# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
# histograms
dataset.hist()
plt.show()
# scatter plot matrix
scatter_matrix(dataset)
plt.show()
# Split-out validation dataset
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size,
random_state=seed)
# Test options and evaluation metric
seed = 7
scoring = 'accuracy'
# Spot Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
# Compare Algorithms
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
# Make predictions on validation dataset
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)
print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))

```



Jason Brownlee March 14, 2018 at 3:10 pm #

REPLY ↩

Perhaps try and run from the command line, not an editor. The editor or notebook can hide output messages and error messages.



Qasem March 14, 2018 at 9:11 pm #

i have solved the problem, where i should close the figures and the results will be displayed, I have tried to change the dataset for example to Heart Dataset, where there are 14 attributes and only two classes, for sure there were an errors. Sir, if I use the heart dataset in which part of the project should I do the modifications? thanks in advance I'm just started to learn Python in Machine learning. your help is really appreciated



Jason Brownlee March 15, 2018 at 6:30 am #

This process will help you work through your problem systematically:
<https://machinelearningmastery.com/start-here/#process>



Daniel March 13, 2018 at 10:50 pm #

REPLY ↩

Jason,

Thanks a bunch for the awesome example. Like others I received 0.991667 for SVM.
The problem, however, I am having relates to the last step – getting prediction values. Below you can find my stack trace.

NOTE: I am mac with python 2.7

Any clue?

ValueError Traceback (most recent call last)

in ()

3 knn.fit(X_train, Y_train)

4 predictions = knn.predict(X_validation)

—> 5 print(accuracy_score(Y_validation, predictions))

6 print(confusion_matrix(Y_validation, predictions))

7 print(classification_report(Y_validation, predictions))

/usr/local/lib/python2.7/site-packages/sklearn/metrics/classification.py in accuracy_score(y_true, y_pred, normalize, sample_weight)
174

175 # Compute accuracy for each possible representation

-> 176 y_type, y_true, y_pred = _check_targets(y_true, y_pred)

177 if y_type.startswith('multilabel'):

178 differing_labels = count_nonzero(y_true - y_pred, axis=1)

/usr/local/lib/python2.7/site-packages/sklearn/metrics/classification.py in _check_targets(y_true, y_pred)

69 y_pred : array or indicator matrix

70 """

—> 71 check_consistent_length(y_true, y_pred)

72 type_true = type_of_target(y_true)

73 type_pred = type_of_target(y_pred)

/usr/local/lib/python2.7/site-packages/sklearn/utils/validation.py in check_consistent_length(*arrays)

202 if len(uniques) > 1:

203 raise ValueError("Found input variables with inconsistent numbers of"

-> 204 " samples: %r" % [int(l) for l in lengths])

205

206

ValueError: Found input variables with inconsistent numbers of samples: [4, 30]



Jason Brownlee March 14, 2018 at 6:23 am #

REPLY ↩

I have not seen this error sorry. Perhaps double check that you have copied all of the code?



Daniel March 16, 2018 at 2:06 am #

REPLY ↩

Found it!!!

Did trv to make some changes in the code but foroot to reverted it back 🙄

...try to make some changes in the code but I get to the same result.

Thanks a lot. That is an awesome example!



Jason Brownlee March 16, 2018 at 6:20 am #

REPLY ↩

Glad to hear it Daniel.



Frank984 March 14, 2018 at 7:46 pm #

REPLY ↩

Hi Jason,

I have a dataset structured as reported here:

<https://app.box.com/s/mi97crz44bz2r7f96wy2z6ztf68ohm87>

(you can download it here: <https://app.box.com/s/c2bxylfe2ggibledjncui05gez13thuo>)

It is composed by 9871 rows e 5 columns:

<https://app.box.com/s/xasyqbhtsmov9gqnv7siop470pgpvvg>

When I try to describe it only the first and second column are considered:

<https://app.box.com/s/9wez8izysrfwivns0sus6ql2ahkq3jc1>

Also if I try to plot a scatter matrix, the data of the first and second column are considered:

<https://app.box.com/s/41x56gxd5bil0c4e0tz000433phoho2v>



Jason Brownlee March 15, 2018 at 6:27 am #

REPLY ↩

Nice work. Note none of your links work.



Frank984 March 15, 2018 at 6:07 pm #

REPLY ↩

I have solved the issue and cancelled the folder.



Jason Brownlee March 16, 2018 at 6:11 am #

REPLY ↩

Great!



Abhay Sapru March 16, 2018 at 6:42 am #

REPLY ↩

till step 5.2 its fine for me but from point 5.3 am getting error as below:-

```
# Spot Check Algorithms
```

```
... models = []
```

```
>>> models.append(('LR', LogisticRegression()))
```

```
Traceback (most recent call last):
```

```
File "", line 1, in
```

```
NameError: name 'LogisticRegression' is not defined
```

```
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
```

```
Traceback (most recent call last):
```

```
File "", line 1, in
```

```
NameError: name 'LinearDiscriminantAnalysis' is not defined
```

```
>>> models.append(('KNN', KNeighborsClassifier()))
```

```
Traceback (most recent call last):
```

```
File "", line 1, in
```

```
NameError: name 'KNeighborsClassifier' is not defined
```

```
>>> models.append(('CART', DecisionTreeClassifier()))
```

```
Traceback (most recent call last):
```

```
File "", line 1, in
```

```
NameError: name 'DecisionTreeClassifier' is not defined
```

```

NameError: name 'DecisionTreeClassifier' is not defined
>>> models.append(('NB', GaussianNB()))
Traceback (most recent call last):
File "", line 1, in
NameError: name 'GaussianNB' is not defined
>>> models.append(('SVM', SVC()))
Traceback (most recent call last):
File "", line 1, in
NameError: name 'SVC' is not defined
>>> # evaluate each model in turn
... results = []
>>> names = []
>>> for name, model in models:
... kfold = model_selection.KFold(n_splits=10, random_state=seed)
File "", line 2
kfold = model_selection.KFold(n_splits=10, random_state=seed)
^
IndentationError: expected an indented block
>>> cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'model_selection' is not defined
>>> results.append(cv_results)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'cv_results' is not defined
>>> names.append(name)
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
Traceback (most recent call last):
File "", line 1, in
NameError: name 'name' is not defined
>>> print(msg)

```



Jason Brownlee March 16, 2018 at 2:20 pm #

REPLY ↩

It looks like you are not preserving the indenting of the code. White space is important in python, the tabs and new lines must be preserved.



Abhay Sapru March 17, 2018 at 8:02 pm #

REPLY ↩

ok i'll try it on ipython may be directly copy paste into command line might have done this and one more thing do i have to define algo names in square brackets and define the seed values in results square brackets



Abhay Sapru March 17, 2018 at 9:56 pm #

REPLY ↩

Below is the code i am trying to run:-

```

1 # Load libraries
2 import pandas
3 from pandas.plotting import scatter_matrix
4 import matplotlib.pyplot as plt
5 from sklearn import model_selection
6 from sklearn.metrics import classification_report
7 from sklearn.metrics import confusion_matrix
8 from sklearn.metrics import accuracy_score
9 from sklearn.linear_model import LogisticRegression
10 from sklearn.tree import DecisionTreeClassifier
11 from sklearn.neighbors import KNeighborsClassifier
12 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
13 from sklearn.naive_bayes import GaussianNB
14 from sklearn.svm import SVC
15 # Load dataset

```

```

16 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
17 names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
18 dataset = pandas.read_csv(url, names=names)
19 # shape
20 print(dataset.shape)
21 # head
22 print(dataset.head(20))
23 # descriptions
24 print(dataset.describe())
25 # class distribution
26 print(dataset.groupby('class').size())
27 # box and whisker plots
28 dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
29 plt.show()
30 # histograms
31 dataset.hist()
32 plt.show()
33 # scatter plot matrix
34 scatter_matrix(dataset)
35 plt.show()
36 # Split-out validation dataset
37 array = dataset.values
38 X = array[:,0:4]
39 Y = array[:,4]
40 validation_size = 0.20
41 seed = 7
42 X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
43 # Test options and evaluation metric
44 seed = 7
45 scoring = 'accuracy'
46 # Spot Check Algorithms
47 models = []
48 models.append(('LR', LogisticRegression()))
49 models.append(('LDA', LinearDiscriminantAnalysis()))
50 models.append(('KNN', KNeighborsClassifier()))
51 models.append(('CART', DecisionTreeClassifier()))
52 models.append(('NB', GaussianNB()))
53 models.append(('SVM', SVC()))
54 # evaluate each model in turn
55 results = []
56 names = []
57 for name, model in models:
58     kfold = model_selection.KFold(n_splits=10, random_state=seed)
59     cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
60     results.append(cv_results)
61     names.append(name)
62     msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
63     print(msg)

```



Katti March 17, 2018 at 2:59 am #

REPLY ↩

Where can we see the visual representation of variate and univariate plots? I'm only seeing textual representation of the data. Please notify where to type dataset.plot(.. code



Katti March 17, 2018 at 3:08 am #

REPLY ↩

My bad,I never used the plt.show() function to visualize my data. I can see the plots very nicely.



Jason Brownlee March 17, 2018 at 8:44 am #

REPLY ↩

Perhaps it would help you to re-read section 4 of the above tutorial?



German Loiti Azcue March 19, 2018 at 8:29 pm #

REPLY ↩

Hi Jason, I really found your guide useful and easy to follow. I am developing my Master Thesis and I am trying to apply ML to predict electricity prices (therefore numerical class). Which algorithm would you recommend me more (more than one if it is possible)?

As far as I know, classification algorithms are used in those cases where the class is binary like in this example. Why do we compare regression model with other classification models in this example then? Does that make sense? Can regression models be applied for classification purposes and vice versa?

Again thanks for your help and your time.



Jason Brownlee March 20, 2018 at 6:17 am #

REPLY ↩

If you are predicting a quantity, you will want to use regression algorithms. I would recommend testing a suite of methods to see which works best on your specific dataset.

Here is more info on the difference between regression and classification:

<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>



Sirish March 22, 2018 at 3:24 am #

REPLY ↩

Why is that same dataset gave two different best machine learning models using two different tools, LDA with R and KNN with Python?



Jason Brownlee March 22, 2018 at 6:26 am #

REPLY ↩

What do you mean exactly?



Vaibhav V March 26, 2018 at 8:56 pm #

REPLY ↩

Well explained concept. Kudos to you.



Jason Brownlee March 27, 2018 at 6:34 am #

REPLY ↩

Thanks!



Danish bhatia March 26, 2018 at 9:18 pm #

REPLY ↩

What is "seed" ?



Jason Brownlee March 27, 2018 at 6:35 am #

REPLY ↩

Good question.

The random number generator used in the splitting of data and within some of the algorithms is actually a pseudorandom number generator. We can seed it so that it will generate the same sequence of random numbers each time the code is run. This helps in tutorials so that you can get the same results that I got.

Learn more about this here:

<https://machinelearningmastery.com/randomness-in-machine-learning/>



Mathew March 27, 2018 at 7:33 am #

REPLY ↩

Hi Jason,

Thank you for the explanation. please find the below questions

1. I changed file name to iris22==> it gave error OK
2. I removed all data in iris.data ==> it gave the same output.
3. If any changes in the iris.data file does not change the output

Can you please explain.

Mathews



Jason Brownlee March 27, 2018 at 4:15 pm #

REPLY ↩

Perhaps confirm that your modified file is still being loaded and used in the code?



Saumya Gupta March 27, 2018 at 10:12 pm #

REPLY ↩

Hey Jason,

I trained my data on a linear regression model, now I want to predict the value of label based on the values of indicators that the user inputs. Can this be done?

I'm really not getting it anywhere.

Please help me out



Jason Brownlee March 28, 2018 at 6:27 am #

REPLY ↩

Linear regression is a model for predicting a quantity, not a label.

This post might clear things up for you:

<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>

Leave a Reply

Name (required)

Email (will not be published) (required)

Website

SUBMIT COMMENT

Welcome to Machine Learning Mastery



Hi, I'm Jason Brownlee, Ph.D.

My goal is to make practitioners like YOU awesome at applied machine learning.

[Read More](#)

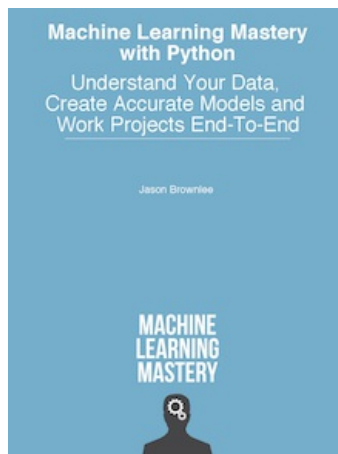
Develop Predictive Models With Python

Want to develop your own models in scikit-learn?

Want step-by-step tutorials?

Looking for sample code and templates?

[Get Started With Machine Learning in Python Today!](#)



POPULAR



Your First Machine Learning Project in Python Step-By-Step

JUNE 10, 2016



Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras

JULY 21, 2016



Multivariate Time Series Forecasting with LSTMs in Keras

AUGUST 14, 2017



How to Setup a Python Environment for Machine Learning and Deep Learning with Anaconda

MARCH 13, 2017



Develop Your First Neural Network in Python With Keras Step-By-Step

MAY 24, 2016



Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras

JULY 26, 2016



Time Series Forecasting with the Long Short-Term Memory Network in Python

APRIL 7, 2017



Regression Tutorial with the Keras Deep Learning Library in Python

JUNE 9, 2016



Multi-Class Classification Tutorial with the Keras Deep Learning Library

JUNE 2, 2016



How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras

AUGUST 9, 2016

Tutorials to Your Inbox

Discover the latest tutorials
in this weekly machine learning newsletter.

Email:

SIGN UP

