☰ | **Navigation**

Want help with Python and scikit-learn? Take the FREE Mini-Course

Search... 🔍

Save and Load Machine Learning Models in Python with scikit-learn

by **Jason Brownlee** on June 8, 2016 in **Python Machine Learning**

Tweet | Share | Share

Finding an accurate machine learning model is not the end of the project.

In this post you will discover how to save and load your machine learning model in Python using scikit-learn.

This allows you to save your model to file and load it later in order to make predictions.

Let's get started.

- **Update Jan/2017**: Updated to reflect changes to the scikit-learn API in version 0.18.
- **Update March/2018**: Added alternate link to download the dataset as the original appears to have been taken down.

Save and Load Machine Learning
Photo by Christine, s

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

[Email Address]

START MY EMAIL COURSE

Need help with Machin

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

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

Start Your FREE Mini-Course Now!

Finalize Your Model with pickle

Pickle is the standard way of serializing objects in Python.

You can use the pickle operation to serialize your machine learning algorithms and save the serialized format to a file.

Later you can load this file to deserialize your model and use it to make new predictions.

The example below demonstrates how you can train a logistic regression model on the Pima Indians onset of diabetes dataset, save the model to file and load it to make predictions on the unseen test set (update: download from here).

```
1   # Save Model Using Pickle
```

```
 2  import pandas
 3  from sklearn import model_selection
 4  from sklearn.linear_model import LogisticRegression
 5  import pickle
 6  url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.c
 7  names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
 8  dataframe = pandas.read_csv(url, names=names)
 9  array = dataframe.values
10  X = array[:,0:8]
11  Y = array[:,8]
12  test_size = 0.33
13  seed = 7
14  X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y, test_size=test_size,
15  # Fit the model on 33%
16  model = LogisticRegression()
17  model.fit(X_train, Y_train)
18  # save the model to disk
19  filename = 'finalized_model.sav'
20  pickle.dump(model, open(filename, 'wb'))
21
22  # some time later...
23
24  # load the model from disk
25  loaded_model = pickle.load(open(filename, 'rb'
26  result = loaded_model.score(X_test, Y_test)
27  print(result)
```

Running the example saves the model to **finalized_m**
saved model and evaluating it provides an estimate o

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
 1  0.755905511811
```

Finalize Your Model with joblib

Joblib is part of the SciPy ecosystem and provides utilities for pipelining Python jobs.

It provides utilities for saving and loading Python objects that make use of NumPy data structures, efficiently.

This can be useful for some machine learning algorithms that require a lot of parameters or store the entire dataset (like K-Nearest Neighbors).

The example below demonstrates how you can train a logistic regression model on the Pima Indians onset of diabetes dataset, saves the model to file using joblib and load it to make predictions on the unseen test set.

```
 1  # Save Model Using joblib
 2  import pandas
 3  from sklearn import model_selection
 4  from sklearn.linear_model import LogisticRegression
 5  from sklearn.externals import joblib
 6  url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.c
 7  names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
 8  dataframe = pandas.read_csv(url, names=names)
 9  array = dataframe.values
10  X = array[:,0:8]
11  Y = array[:,8]
12  test_size = 0.33
```

```
13  seed = 7
14  X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y, test_size=test_size,
15  # Fit the model on 33%
16  model = LogisticRegression()
17  model.fit(X_train, Y_train)
18  # save the model to disk
19  filename = 'finalized_model.sav'
20  joblib.dump(model, filename)
21
22  # some time later...
23
24  # load the model from disk
25  loaded_model = joblib.load(filename)
26  result = loaded_model.score(X_test, Y_test)
27  print(result)
```

Running the example saves the model to file as **finalized_model.sav** and also creates one file for each NumPy array in the model (four additional files). After the model on unseen data is reported.

```
1  0.755905511811
```

Tips for Finalizing Your Model

This section lists some important considerations when

- **Python Version**. Take note of the python version.
  maybe minor) version of Python used to serialize
- **Library Versions**. The version of all major libraries
  certainly need to be the same when deserializing a saved model. This is not limited to the version of
  NumPy and the version of scikit-learn.
- **Manual Serialization**. You might like to manually output the parameters of your learned model so that
  you can use them directly in scikit-learn or another platform in the future. Often the algorithms used
  by machine learning algorithms to make predictions are a lot simpler than those used to learn the
  parameters can may be easy to implement in custom code that you have control over.

Take note of the version so that you can re-create the environment if for some reason you cannot reload
your model on another machine or another platform at a later time.

Summary

In this post you discovered how to persist your machine learning algorithms in Python with scikit-learn.

You learned two techniques that you can use:

- The pickle API for serializing standard Python objects.
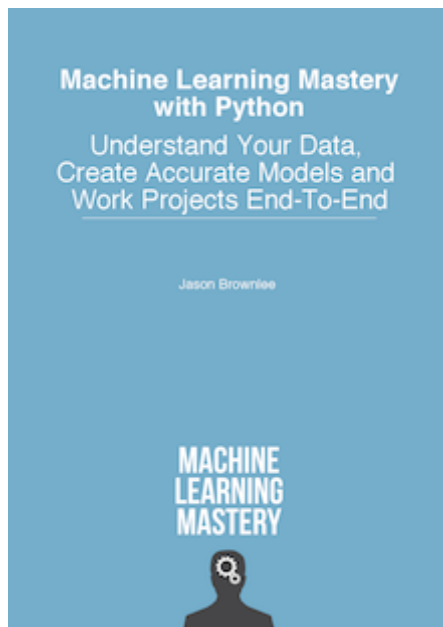- The joblib API for efficiently serializing Python objects with NumPy arrays.

Do you have any questions about saving and loading your machine learning algorithms or about this post?
Ask your questions in the comments and I will do my best to answer them.

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…

Fir

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Tweet   Share   Share

About Jason Brownlee

Jason Brownlee, PhD 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 →

‹ Binary Classification Tutorial with the Keras Deep Learning Library

Regression Tutorial with the Keras Deep Learning Library in Python ›

180 Responses to *Save and Load Machine Learning Models in Python with scikit-learn*

Kayode October 18, 2016 at 6:15 pm #   REPLY ↰

Thank you so much for this educative post.

Jason Brownlee October 19, 2016 at 9:17 am #

You're welcome Kayode.

TonyD November 13, 2016 at 3:52 pm #

Hi Jason,

I have two of your books and they are awesome. I took several machine learning courses before, however as you mentioned they are more geared towards theory than practicing. I devoured your Machine Learnign with Python book and 20x my skills compared to the c

I found this page by Googling a code snippet in chapt

loaded_model = pickle.load(open(filename, 'rb'))

throws the error:

runfile('C:/Users/Tony/Documents/MassData_Regressi
File "C:/Users/Tony/Documents/MassData_Regression
loaded_model = pickle.load(open(filename, 'rb'))
^
SyntaxError: invalid syntax

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

> Email Address

START MY EMAIL COURSE

Jason Brownlee November 14, 2016 at 7:36 am #

Thanks TonyD.

I wonder if there is a copy-paste error, like an extra space or something?

Does the code example (.py file) provided with the book for that chapter work for you?

William January 7, 2019 at 9:37 pm #

As Jason already said, this is a copy paste problem. In your line specifically, the quotes are the problem.

```
loaded_model = pickle.load(open(filename, 'rb'))
```

It should be

```
loaded_model = pickle.load(open(filename, 'rb'))
```

Try to understand the difference :).

Jason Brownlee January 8, 2019 at 6:49 am #   REPLY ↩

Thanks.

This might help:

https://machinelearningmastery.com/faq/single-faq/how-do-i-copy-code-from-a-tutorial

Konstantin November 19, 2016 at 6:01 am #   REPLY ↩

Hello, Jason

Where we can get X_test, Y_test "sometime later"? It i
X_test, Y_test not pickled In your example you pickle
applications is not single flow I found work around and

What is correct solution? Should we pickle decorator
values? I didn't find legal information from documenta
to pull Y values from classifier.

Can you advise?

> ## Your Start in Machine Learning   ✕
>
> You can master applied Machine Learning
> **without math or fancy degrees**.
> Find out how in this *free* and *practical* course.
>
> Email Address
>
> **START MY EMAIL COURSE**

Jason Brownlee November 19, 2016 at 8:51 a

Hi Konstantin,

I would not suggest saving the data. The idea is to show how to load the model and use it on new data
– I use existing data just for demonstration purposes.

You can load new data from file in the future when you load your model and use that new data to make
a prediction.

If you have the expected values also (y), you can compare the predictions to the expected values and
see how well the model performed.

Guangping Zhang November 21, 2016 at 6:01 am #   REPLY ↩

I'm newer Pythoner, your code works perfect! But where is the saved file? I used windows
10.

Jason Brownlee November 22, 2016 at 6:56 am #   REPLY ↩

Thanks Guangping.

The save file is in your current working directory, when running from the commandline.

If you're using a notebook or IDE, I don't know where the file is placed.

Mohammed Alnemari December 13, 2016 at 2:45 pm #

Hi Jason ,

I am just wondering if can we use Yaml or Json with sklearn library . I tried to do it many times but I could not reach to an answer . I tried to do it as your lesson of Kares , but for some reason is not working . hopefully you can help me if it is possible

Jason Brownlee December 14, 2016 at 8:24 a

Hi Mohammed, I believe the serialization
library.

sklearn serialization is focused on binary files like

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Normando Zubia December 29, 2016 at 9:55 am #

Hi, my name is Normando Zubia and I have b

lessons.

I'm currently working on a model to predict user behavoir in a production environment. Due to several situations I can not save the model in a pickle file. Do you know any way to save the model in a json file?

I have been playing a little with sklearn classes and I noticed that if I save some parameters for example: n_values_, feature_indices_ and active_features_ in a OneHotEncoding model I can reproduce the results. Could this be done with a pipeline? Or do you think I need to save each model's parameters to load each model?

PS: Sorry for my bad english and thanks for your attention.

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

Hi Normando,

If you are using a simple model, you could save the coefficients directly to file. You can then try and put them back in a new model later or implement the prediction part of the algorithm yourself (very easy for most methods).

Let me know how you go.

Samuel February 6, 2017 at 3:14 pm #

Hello Jason,

I am new to machine learning. I am your big fan and read a lot of your blog and books. Thank you very much for teaching us machine learning.

I tried to pickle my model but fail. My model is using VGG16 and replace the top layer for my classification solution. I further narrowed down the problem and find that it is the VGG16 model failed to pickle. Please find my simplified code below and error log below:

It will be highly appreciated if you can give me some direction on how to fix this error.

Thank you very much
————————————————-
# Save Model Using Pickle
from keras.applications.vgg16 import VGG16
import pickle

model = VGG16(weights='imagenet', include_top=Fals

filename = 'finalized_model.sav'
pickle.dump(model, open(filename, 'wb'))

————————————————-

/Library/Frameworks/Python.framework/Versions/2.7/k
/Users/samueltin/Projects/bitbucket/share-card-ml/pic
Using TensorFlow backend.
Traceback (most recent call last):
File "/Users/samueltin/Projects/bitbucket/share-card-ml/pickle_test.py", line 8, in
pickle.dump(model, open(filename, 'wb'))
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 1376, in dump
Pickler(file, protocol).dump(obj)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 224, in dump
self.save(obj)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save
self.save_reduce(obj=obj, *rv)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce
save(state)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 606, in save_list

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
self._batch_appends(iter(obj))
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 621, in
_batch_appends
save(x)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save
self.save_reduce(obj=obj, *rv)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce
save(state)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions
self._batch_appends(iter(obj))
File "/Library/Frameworks/Python.framework/Versions
_batch_appends
save(x)
File "/Library/Frameworks/Python.framework/Versions
self.save_reduce(obj=obj, *rv)
File "/Library/Frameworks/Python.framework/Versions
save(state)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save
self.save_reduce(obj=obj, *rv)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce
save(state)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save
self.save_reduce(obj=obj, *rv)
```

Your Start in Machine Learning

×

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce

save(state)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save

f(self, obj) # Call unbound method with explicit self

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict

self._batch_setitems(obj.iteritems())

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in

_batch_setitems

save(v)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save

f(self, obj) # Call unbound method with explicit self

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict

self._batch_setitems(obj.iteritems())

File "/Library/Frameworks/Python.framework/Versions/_batch_setitems

save(v)

File "/Library/Frameworks/Python.framework/Versions/f(self, obj) # Call unbound method with explicit self

File "/Library/Frameworks/Python.framework/Versions/self._batch_appends(iter(obj))

File "/Library/Frameworks/Python.framework/Versions/_batch_appends

save(x)

File "/Library/Frameworks/Python.framework/Versions/self.save_reduce(obj=obj, *rv)

Your Start in Machine Learning

×

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce

save(state)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save

f(self, obj) # Call unbound method with explicit self

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict

self._batch_setitems(obj.iteritems())

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in

_batch_setitems

save(v)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save

self.save_reduce(obj=obj, *rv)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce

save(state)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save

f(self, obj) # Call unbound method with explicit self

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict

self._batch_setitems(obj.iteritems())

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in

_batch_setitems

save(v)

File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 331, in save

```
self.save_reduce(obj=obj, *rv)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 425, in save_reduce
save(state)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 655, in save_dict
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 286, in save
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/...
self._batch_appends(iter(obj))
File "/Library/Frameworks/Python.framework/Versions/...
_batch_appends
save(x)
File "/Library/Frameworks/Python.framework/Versions/...
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/...
save(element)
File "/Library/Frameworks/Python.framework/Versions/...
f(self, obj) # Call unbound method with explicit self
File "/Library/Frameworks/Python.framework/Versions/...
self._batch_setitems(obj.iteritems())
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 669, in
_batch_setitems
save(v)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/pickle.py", line 306, in save
rv = reduce(self.proto)
File "/Library/Frameworks/Python.framework/Versions/2.7/lib/python2.7/copy_reg.py", line 70, in
_reduce_ex
raise TypeError, "can't pickle %s objects" % base.__name__
TypeError: can't pickle module objects

Process finished with exit code 1
```

Jason Brownlee February 7, 2017 at 10:11 am #   REPLY ↰

Sorry Samuel, I have not tried to save a pre-trained model before. I don't have good advice for
you.

Let me know how you go.

huikang September 21, 2018 at 11:50 am #

Is there a more efficient method in machine learning than joblib.load(), storing the model directly in memory and using it again?

Jason Brownlee September 21, 2018 at 2:21 pm #

Sure, you can make an in-memory copy. I think sklearn has a clone() function that you can use.

Amy March 8, 2017 at 7:03 am #

I have trained a model using liblinearutils. The error that ctype module with pointers cannot be pickle

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee March 8, 2017 at 9:47 am #

Sorry Amy, I don't have any specific exam

Perhaps you can save the coefficients of your mo

SHUBHAM BHARDWAJ April 3, 2017 at 10:42 pm #

Thanks a lot, very useful

Jason Brownlee April 4, 2017 at 9:14 am #

You're welcome!

Benju April 11, 2017 at 1:35 am #

My saved modells are 500MB+ Big….is that normal?

Jason Brownlee April 11, 2017 at 9:34 am #

Ouch, that does sound big.

If your model is large (lots of layers and neurons) then this may make sense.

Anupam April 13, 2017 at 2:32 am #   REPLY ↩

How to use model file ("finalized_model.sav") to test unknown data. Like, if the model is for tagger , how this model will tag the text file data? Is there any example?

Jason Brownlee April 13, 2017 at 10:08 am #

You can load the saved model and start n

See this post on finalizing models:

http://machinelearningmastery.com/train-final-mad

Your Start in Machine Learning   ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Oss Mps April 21, 2017 at 3:09 pm #

Dear Sir, please advice on how to extract wei

Email Address

START MY EMAIL COURSE

Jason Brownlee April 22, 2017 at 9:23 am #   REPLY ↩

I would suggest extracting coefficients from your model directly and saving them in your preferred format.

Suhas May 24, 2017 at 4:44 am #   REPLY ↩

Hi I love your website; it's very useful!

Are there any examples showing how to save out the training of a model after say 100 epochs/iterations? It's not immediately clear from looking at joblib or scikit learn.

This is esp. useful when dealing with large datasets and/or computers or clusters which may be unreliable (e.g., subject to system reboots, etc.)

Jason Brownlee May 24, 2017 at 4:59 am #   REPLY ↩

I'm not sure how to do this with sklearn. You may need to write something custom. Consider posting to stackoverflow.

Viktor May 30, 2017 at 8:52 am #   REPLY ↩

Hey!
Is it possible to open my saved model and make a prediction on cloud server where is no sklearn installed?

Jason Brownlee June 2, 2017 at 12:31 pm #   REPLY ↩

no.

You could save the coefficients from within the mo...
code.

Your Start in Machine Learning   ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Clemence June 8, 2017 at 6:55 pm #

Hello Jason and thank you very much, it's bee...

Do you know if it's possible to load features transform...
I'm mostly thinking of categorical variables that we ne...

I'm using sklearn to do that, but I don't know if we can (as for Spark), integrate this transformation with the ML model into the serialized file (Pickle or Joblib).

#Encode categorical variable into numerical ones
from sklearn.preprocessing import LabelEncoder
list_var = ['country', 'city']

encoder = LabelEncoder()
for i in list_var:
df[i] = encoder.fit_transform(df[i])

Then I fit the model on the training dataset...

And I need to save this transformation with the model. Do you know if that's possible ?
Thank you!

Jason Brownlee June 9, 2017 at 6:23 am #   REPLY ↩

I'm not sure I follow sorry.

You can transform your data for your model, and you can apply this same transform in the future when you load your model.

You can save the transform objects using pickle. Is that what you mean?

Bhavani Shanker June 22, 2017 at 1:24 am #

Hi Jason,

Kindly accept my encomiums for the illustrative lecture that you have delivered on Machine Learning using Python.

```
**********************************************
# save the model to disk
filename = 'finalized_model.sav'
joblib.dump(model, filename)

# sometime later…

# load the model from disk
loaded_model = joblib.load(filename)
result = loaded_model.score(X_test, Y_test)
print(result)
*********************************************************
```

After saving the model 'finalized_model.sav' , How can ~~~~ date?

I would appreciate if you can advice on this

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee June 22, 2017 at 6:11 am #

The code after "sometime later" would be in a new session.

---

jinsh June 28, 2017 at 8:57 pm #

Hello sir,

The above code saves the model and later we can check the accuracy also
but what i have to do for making predicting the class of unknown data?
I mean which function have to be called ?

eg: 2,132,40,35,168,43.1,2.288,33

can you suggest how to get the class of above data through prediction ?

thank you

Jason Brownlee June 29, 2017 at 6:35 am #

Pass in input data to the predict function and use the result.

```
1  yhat = model.predict(X)
```

Ukesh Chawal July 24, 2017 at 11:09 pm #

Can we use "pickling" to save an LSTM model and to load or used a hard-coded pre-fit model to generate forecasts based on data passed in to initialize the model?

When I tried to use it, it gave me following error:

PicklingError: Can't pickle : attribute lookup module or

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee July 25, 2017 at 9:44 am #

No.

See this tutorial on how to save Keras models:

http://machinelearningmastery.com/save-load-ker

Ukesh Chawal July 25, 2017 at 11:59 pm #

Great. It worked.

You are awesome Jason. Appreciated.

Jason Brownlee July 26, 2017 at 7:57 am #

Glad to hear it.

akatsuki August 9, 2017 at 1:21 pm #

tbh this is best of the sites on web. Great!
I love the email subscriptions of yours as a beginner they are quite helpful to me .

Jason Brownlee August 10, 2017 at 6:45 am #

Thanks, I'm glad to hear that.

vikash August 10, 2017 at 9:32 pm #

Hi @Jason Brownlee thanks for such informative blog. Can you please guide me for a problem where i would like to retrain the .pkl model only with new dataset with new class keeping the previous learning intact. I had thought that model.fit(dataset,label) will do that but it forgets the previous learning. Please suggest me some techniques for it.
Thanks

Jason Brownlee August 11, 2017 at 6:42 am

Sorry, I don't follow. Can you please resta

Your Start in Machine Learning   ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

sassashi August 28, 2017 at 4:41 am #

Hi Jason, I believe @vikash is looking
examples after the initial training stage. This is
possible to retrain a model in tensorflow with n
sklearn.

to expand the question some more: 1-you train a model with sklearn 2-save it with pickle or joblib 3-then you get your hands on some new examples that were not available at the time of initial training "step 1" 4-you load the previous model 5-and now you try to train the model again using the new data without losing the previous knowledge… is step 5 possible with sklearn?

Jason Brownlee August 28, 2017 at 6:52 am #

I have not updated a model in sklearn, but I would expect you can.

Here is an example of updating a model in Keras which may help in general principle:
https://machinelearningmastery.com/update-lstm-networks-training-time-series-forecasting/

Navdeep Singh August 22, 2017 at 8:30 pm #

Hi Json,

I need your guidance on Updation of saved pickle files with new data coming in for training

I recall 3 methods, Online Learning which is train one every new observation coming in and in this case model would always be biased towards new features ,which i dont wana do

Second is, Whenever some set of n observations comes, embedd it with previous data and do retraining again from scratch, that i dont want to do as in live environment it will take lot of time

Third is Mini batch learning, i know some of algorithm like SGD and other use partial fit method and do same but I have other algorithms as week like random forest , decision tress, logistic regression. I wana ask can i update the previously trained pickle with new training ?

I am doing it in text classification, I read that possibly doing this, model update pickle will not take new features of new data (made using tfidf or countvectorizer) and it would be of less help.

Also as domain is same, and If client(Project we are working for) is different , inspite of sharing old data with new client (new project), could i use old client trained r̶ ⬚ data. Basically I am transferring learning

Your Start in Machine Learning ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee August 23, 2017 at 6:48 am

Great question.

This is a challenging problem to solve. Really, the ⬚
requirements.

A flexible approach may be to build-in capacity int ⬚
future.

The simplest approach is to ignore new words.

These, and other strategies are testable. See how performance degrades under both schemes with out-of-band test data.

Merari September 11, 2017 at 7:59 am #   REPLY ↩

Gracias por compartir,
Existe alguna forma en la que pueda realizar predicciones con nuevos datos solo con el modelo guardado? llamando este modelo desde un archivo nuevo? lo he intentado con la instruccion final:

```
# load the model from disk
loaded_model = pickle.load(open(filename, 'rb'))
result = loaded_model.score(X_test, Y_test)
print(result)
```

pero no lo he logrado

373/5000
Thanks for sharing,
Is there any way I can make predictions with new data only with the saved model? calling this model from a new file? I have tried with the final instruction:

```
# load the model from disk
loaded_model = pickle.load (open (filename, 'rb'))
result = loaded_model.score (X_test, Y_test)
print (result)
```

but I have not achieved it

Jason Brownlee September 11, 2017 at 12:11 pm #   REPLY ↩

That is exactly what we do in this tutorial.

What is the problem exactly?

AP September 29, 2017 at 6:36 am #

Hi Jason, I learn a lot reading your python bo

I'm having an issue when I work on text data with load
training data with countvectorizer and tfidf. Then I only
usual. But, when work on loaded pretrained model in a
extraction. I can't just transform the test data as it ask
session. If I fit and transform on test data only, model
believe that is wrong way of doing machine learning. S
countvectorizer, tfidf or other cases while working with previously trained model?

I'm using spark ML but I think it would be the same for scikit-learn as well.

> ## Your Start in Machine Learning   ✕
>
> You can master applied Machine Learning
> **without math or fancy degrees**.
> Find out how in this *free* and *practical* course.
>
> [Email Address   ]
>
> **START MY EMAIL COURSE**

Jason Brownlee September 30, 2017 at 7:31 am #   REPLY ↩

Perhaps you can pickle your data transform objects as well, and re-use them in the second
session?

Bhavya Chugh October 29, 2017 at 5:57 am #   REPLY ↩

Hi Jason,

I trained a random forest model and saved the same as a pickle file in my local desktop. I then copied that
pickle file to my remote and tested the model with the same file and it is giving incorrect predictions. I am
using python 3.6 in my local and python 3.4 in my remote, however the version of scikit-learn are same. Any
ideas why this may be happening?

Jason Brownlee October 29, 2017 at 6:00 am # REPLY ↩

No idea, perhaps see if the experiment can be replicated on the same machine? or different machines with the same version of Python?

Berkin Albert Antony November 10, 2017 at 5:45 pm # REPLY ↩

Hi Jason Brownlee,

I have a LogisticRegression model for binary classification. I wish to find a similar data points in a trained model for a given test data points. So that I can show these are similar data points predicted with these same class.

Could you please suggest your thoughts for the same.

Thanks

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee November 11, 2017 at 9:18 a

Perhaps you could find data points with a

REPLY ↩

James November 16, 2017 at 8:47 am # REPLY ↩

Hi Jason –

If you pickle a model trained on a subset of features, is it possible to view these features after loading the pickled model in a different file? For example: original df has features a,b,c,d,e,f. You train the model on a,c,e. Is it possible to load the pickled model in a separate script and see the model was trained on a,c,e?

Thanks,
James

Jason Brownlee November 16, 2017 at 10:33 am # REPLY ↩

Yes, you can save your model, load your model, then use it to make predictions on new data.

Mrinal Mitra November 22, 2017 at 6:26 am # REPLY ↩

Hi Jason,

Thanks for explaining it so nicely. I am new to this and will be needing your guidance. I have data using which I have trained the model. Now I want this model to predict an untested data set. However, my requirement is an output which will have the data and corresponding prediction by the model. For example, record 1 – type a, record 2 – type a, record 3 – type c and so on. Could you please guide me on this?

Jason Brownlee November 22, 2017 at 11:16 am #

You can provide predictions one at a time or in a group to the model and the predictions will be in the same order as the inputs.

Does that help?

Niranjan December 3, 2017 at 3:22 pm #

Hi,

I am using chunks functionality in the read csv method
and save it. But it always saves the model that is being
you help me with it

clf_SGD = SGDClassifier(loss='modified_huber', penal
pd.read_csv("file_name",chunksize = 1000):
""""
data preparation and cleaning
""""
hashing = hv.fit_transform(X_train['description'])
clf_SGD.partial_fit(hashing, y_train, classes= y_classes)

joblib.dump(clf_SGD, source_folder + os.path.sep+'text_clf_sgd.pkl')

Jason Brownlee December 4, 2017 at 7:46 am #

Sorry, I'm not sure I follow, could you please try reframing your question?

Shabbir December 8, 2017 at 8:50 am #

Hi Jason,
This is extremely helpful and saved me quite a bit of processing time.

I was training a Random Forest Classifier on a 250MB data which took 40 min to train everytime but results were accurate as required. The joblib method created a 4GB model file but the time was cut down to 7 Minutes to load. That was helpful but the results got inaccurate or atleast varied quite a bit from the original

results. I use average of 2 Decision Tree and 1 Random Forest for the model. Decision Tree Models have kept there consistency loading vs training but RF hasn't. Any ideas?

Nilanka December 19, 2017 at 9:10 pm #

Thank you very useful!!

Jason Brownlee December 20, 2017 at 5:43 am #

You're welcome.

Gokhan December 28, 2017 at 2:55 pm #

Hello, if i load model
loaded_model = joblib.load(filename)
result = loaded_model.score(X_test, Y_test)
print(result)

can i use this model for another testsets to prediction?

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee December 29, 2017 at 5:17 am #

Sure.

Vinay Boddula January 20, 2018 at 5:31 am #

Hi Jason,

How do I generated new X_Test for prediction ? This new X_Test needs to make sure that the passed parameters are same in the model was trained with.

Background: I am basically saving the model and predicting with new values from time to time. How do we check whether the new values have all the parameters and correct data type.

Jason Brownlee January 20, 2018 at 8:25 am #

Visualization and statistics.

I have many posts on the topic, try the search box.

Sekar February 1, 2018 at 4:06 am #   REPLY ↩

Jason. Very good article. As asked by others, in my case I am using DecisionTreeClassifier with text feature to int transformation. Eventhough, you mentioned that transformation map can also be picked and read back, is there any example available? Will it be stored in the same file or it will be another file?

Jason Brownlee February 1, 2018 at 7:24 am   REPLY ↩

In a separate file.

Your Start in Machine Learning   ✕

Yousif February 5, 2018 at 8:01 pm #

Thank you so much professor
we get more new knowledge

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee February 6, 2018 at 9:12 am

You're welcome. Also, I'm not a professor.

Adarsh C February 8, 2018 at 12:29 pm #   REPLY ↩

HI sir,
I would like to save predicted output as a CSV file. After doing ML variable I would like to save
"y_predicted". And I'm using python ide 3.5.x I have pandas,sklearn,tensorflow libraries

---

Jason Brownlee February 9, 2018 at 8:58 am #   REPLY ↩

You can save the numpy array as a csv.
https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.savetxt.html

---

Atul March 11, 2018 at 6:45 am #   REPLY ↩

Hi Jason,

I would like to save predicted output as a CSV file. After doing ML variable I would like to save "y_predicted". How I can save Naive Bayes, SVM, RF and DT Classification for final predictions for all samples saved as a .csv with three columns namely Sample, Actual value, Prediction values

Jason Brownlee March 12, 2018 at 6:24 am #   REPLY ↩

Perhaps create a dataframe with all the columns you require and save the dataframe directly via to_csv():

https://pandas.pydata.org/pandas-docs/stable/ge

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**. Find out how in this *free* and *practical* course.

| Email Address |

START MY EMAIL COURSE

Tommy March 22, 2018 at 11:14 pm #

I have a list of regression coefficients from a p
the sklearn logistic regression function to try and repro
Thanks!
Tommy

Jason Brownlee March 23, 2018 at 6:07 am #   REPLY ↩

No model is needed, use each coefficient to weight the inputs on the data, the weighted sum is the prediction.

Vincent April 10, 2018 at 10:25 am #   REPLY ↩

Hi,all
I am using scikit 0.19.1
I generated a training model using random forest and saved the model. These were done on ubuntu 16.01 x86_64.
I copied the model to a windows 10 64 bit machine and wanted to reuse the saved model. But unfortunately i get the following
Traceback (most recent call last):
File "C:\Users\PC\Documents\Vincent\nicholas\feverwizard.py.py", line 19, in
rfmodel=joblib.load(modelfile)
File "C:\Python27\lib\site-packages\sklearn\externals\joblib\numpy_pickle.py", line 578, in load
obj = _unpickle(fobj, filename, mmap_mode)
File "C:\Python27\lib\site-packages\sklearn\externals\joblib\numpy_pickle.py", line 508, in _unpickle
obj = unpickler.load()

File "C:\Python27\lib\pickle.py", line 864, in load
dispatchkey
File "C:\Python27\lib\pickle.py", line 1139, in load_reduce
value = func(*args)
File "sklearn\tree_tree.pyx", line 601, in sklearn.tree._tree.Tree.cinit
ValueError: Buffer dtype mismatch, expected 'SIZE_t' but got 'long long'

What could be happening? Is it because of a switch from ubuntu to windows? However i am able to reuse the model in my ubuntu.

Jason Brownlee April 11, 2018 at 6:29 am #　　　　　　　　REPLY ↩

Perhaps the pickle file is not portable acr

Pramod April 17, 2018 at 9:03 pm #

Can we load model trained on 64 bit system

Jason Brownlee April 18, 2018 at 8:04 am #

I'm skeptical that it would work. Try it and see. Let me know how you go.

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Arnaud April 17, 2018 at 9:29 pm #　　　　　　　　REPLY ↩

Dear Jason :

Thank you for 'le cours' which is very comprehensive.

I have a maybe tricky but 'could be very usefull' question about my newly created standard Python object.

Is it possible to integrate a call to my Python object in a Fortran program ?

Basically I have a deterministic model in which I would like to make recursive calls to my Python object at every time step.

Do I need some specific libraries ?

Thank you
Best regards

Jason Brownlee April 18, 2018 at 8:06 am #　　　　　　　　REPLY ↩

You're welcome.

I suspect it is possible. It's all just code at the end of the day. You might need some kind of Python-FORTRAN bridge software. I have not done this, sorry.

Pratip April 23, 2018 at 4:32 pm #

Hi Sir ,
I wanted to know if its possible to combine the scikit preloaded datasets with some new datasets to get more training data to get further higher accuracy or firstly run on the scikit loaded dataset and then save model using pickle an run it on another dataset .
Which method will be correct ?
Please help .

Jason Brownlee April 24, 2018 at 6:20 am #

Sure, you can, but it may only make sens~~e~~

same domain.

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

swati kumari April 12, 2019 at 5:07 pm #

How it can be done? When I am loading the pickle and try to fit new data , the model gets fitted with new data only.

Jason Brownlee April 13, 2019 at 6:24 am #

If the model has already been fit, saved, loaded and is then trained on new data, then it is being updated, not trained from scratch.

Perhaps I don't understand the problem you're having?

Ishit Gandhi May 4, 2018 at 6:00 pm #

Hii Jason,

Can you put example of how to store and load Pipeline models?

eg.

```
clf = Pipeline([("rbm",rbm),("logistic",logistic)])
clf.fit(trainX,trainY)
```

Jason Brownlee May 5, 2018 at 6:18 am #   REPLY ↩

Perhaps use pickle? This might help:

https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/

Akash May 14, 2018 at 4:15 pm #   REPLY ↩

Hi jason,
My name is Akash Joshi.I am trying to train my scikit s
memory.Is there a way where I can train the svm mode

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Jason Brownlee May 15, 2018 at 7:51 am #
[Email Address]

Perhaps try running on a machine with m

Perhaps try using a sample of your dataset instead

Perhaps use a generator to progressively load the

START MY EMAIL COURSE

Samarth May 14, 2018 at 4:54 pm #   REPLY ↩

Hi Jason

I want to know how can presist a minmax transjformation? There are ways to persist the final model but to
persist the transformations?

Thanks

Jason Brownlee May 15, 2018 at 7:51 am #   REPLY ↩

Save the min and max values for each variable.

Or save the whole object.

SOORAJ T S May 16, 2018 at 12:30 am #   REPLY ↩

thank you the post, it is very informative but i have a doubt about the labels or names of the dataset can specify each.

Jason Brownlee May 16, 2018 at 6:05 am # REPLY ↩

What do you mean exactly?

SOORAJ T S May 16, 2018 at 4:11 pm # REPLY ↩

names = ['preg', 'plas', 'pres', 'skin', 'test', 'm

in the above code what are these "preg" , "plas", "pres

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Jason Brownlee May 17, 2018 at 6:24 am #

You can learn about these features here:

https://github.com/jbrownlee/Datasets/blob/maste

Email Address

START MY EMAIL COURSE

SOORAJ T S May 17, 2018 at 4:23 pm #

thank you sir…

Aniko June 7, 2018 at 12:13 am # REPLY ↩

HI Jason!

I created a machine learning (GBM) model to predict house prices and a Django application to usability. This model has more than 1000 n_estimators and it takes more than 1 minutes to load before getting the prediction in every request.
I would like to load joblib dump file just once and store the model in memory, avoiding loading the model in every get requests.

What is your best practice for this?

Thanks

Jason Brownlee June 7, 2018 at 6:31 am # REPLY ↩

This sounds like a web application software engineering question rather than a machine learning question.

Perhaps you can host the model behind a web service?

Aniko June 7, 2018 at 6:51 pm #   REPLY ↰

thank you, meanwhile I found some caches -related solution in Django documentation, this perhaps solve the loading problem

---

Jason Brownlee June 8, 2018 at 6:

Glad to hear it.

Your Start in Machine Learning   ✕

LamaOS223 June 9, 2018 at 2:00 pm #

okay what if i had 2 datasets for Example a L   You can master applied Machine Learning
the first dataset has a Loan_Status attribute   **without math or fancy degrees**.
and the second one does not have a Loan_Status attri   Find out how in this *free* and *practical* course.
if i trained the model on the first dataset and i want to
to do that? please make it simple for me i'm beginner

Email Address

START MY EMAIL COURSE

Jason Brownlee June 10, 2018 at 5:58 am #   REPLY ↰

This process wills how you how to work through a predictive model systematically:
https://machinelearningmastery.com/start-here/#process

Imti July 12, 2018 at 4:55 pm #   REPLY ↰

Hey Jason, I am working on a model to classify text files. I am using the CountVectorizer, TfidfTransformer and SGDClassifier in the same sequence on a set of data files. I am saving the SGDClassifier object via the joblib.dump method you have mentioned in this article.

Do I also need to save the vectorizer and transformer objects/models ? Since when i take a new file for classification I will need to go through these steps again.

Jason Brownlee July 13, 2018 at 7:33 am #   REPLY ↰

Yes, they are needed to prepare any data prior to using the model.

Dennis Faucher July 28, 2018 at 2:38 am #   REPLY ↰

Just what I needed today. Thank syou.

Jason Brownlee July 28, 2018 at 6:38 am #   REPLY ↰

I'm happy to hear that Dennis.

Tejaswini July 30, 2018 at 9:01 am #

Hi Jason,
Appreciate for the article. when i am saving the model
different accuracy.

Problem trying to solve: I am using oneclasssvm mode

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee July 30, 2018 at 2:15 pm #

I have not seen that, are you sure you are evaluating the model on exactly the same data?

Tejaswini August 2, 2018 at 2:10 pm #   REPLY ↰

Yes Jason i am using gensim word2vec to convert text into feature vectors and then performing classification task.after saving model and reloading in another session its giving different results.

Jason Brownlee August 2, 2018 at 2:11 pm #   REPLY ↰

That is odd. I have not seen this.

Perhaps report a fault/bug?

EvapStudent August 7, 2018 at 1:36 am #   REPLY ↰

Hi Jason,

I am training a neural network using MLPRegressor, trying to predict pressure drop in different geometries of heat exchangers. I think I have gotten the network to train well with low MRE, but I can't figure out how to use the network. When I tried to load using pickle and call again, I am getting an error when using "score". I am new to python so not sure how to go about bringing in new data for the network to predict or how to generalize doing so.

Jason Brownlee August 7, 2018 at 6:28 am #   REPLY ↩

I don't recommend using pickle. I recommend using the Keras API to save/load your model.

Once you find a config that works for your problem   Keras API directly.

EvapStudent August 7, 2018 at 11:13 pm

Hi Jason,

Thanks for the recommendation. Is there no ea
scikit learn? I have been getting good results v
how to get it to the point where I can actually u
predictions).

If using Keras API to save/load is the best option, how do I go about doing that?

Your Start in Machine Learning   ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee August 8, 2018 at 6:21 am #   REPLY ↩

There may be, but I don't have an example, sorry.

Golnoush August 21, 2018 at 1:38 am #   REPLY ↩

Hello Jason,

Thank you for your nice tutorial! Does pickle.dump(model, open(filename, 'wb')) only save the neural network model or it also save the parameters and weights of the model?
Does the back propagation and training is done again when we use pickle.load ?
What I would like to do is that I aim to save the whole model and weights and parameters during training and use the same trained model for every testing data I have. I would be so thankful if you could assist me in this way.

Jason Brownlee August 21, 2018 at 6:19 am #

I believe you cannot use pickle for neural network models – e.g. Keras models.

Somo August 29, 2018 at 3:05 pm #

Hi Jason,

I am trying to save my model using joblib.dump(model, 'model.pkl') and load it back up in another .py file model = joblib.load('model.pkl') but then the accuracy dropped and each time I run it the accuracy differs a lot. I coefficient and the intercept and the same for both models. Any ideas why this might happen. Thanks in advance.

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Jason Brownlee August 30, 2018 at 6:26 am

Perhaps this will help:
https://machinelearningmastery.com/faq/single-fa
code

Email Address

START MY EMAIL COURSE

Dhrumil September 1, 2018 at 3:11 pm #

Hey man I am facing a trouble with pickle, when I try to load my .pkl model I am getting following error :

UnicodeDecodeError: 'ascii' codec can't decode byte 0xbe in position 3: ordinal not in range(128)

Can you please tell me something since I have tried all fixes I could find..

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

Perhaps post your error on stackoverflow?

Aakash Aggarwal September 8, 2018 at 4:57 am #

I want to develop to train my model and save in pickle file. From the next time onwards, when i want to train the model, it should save in previously created pickle file in append mode that reduces the time of training the model. I am using LogisticRegression model.

Any helps would be greatly appreciated.

Jason Brownlee September 8, 2018 at 6:17 am #

This post shows how:

https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/

Aakash Aggarwal October 2, 2018 at 12:45 am #

This article shows how to save a model that is built from scratch. But I am looking to train the model by including additional data so as to achieve high prediction performance and accuracy for unseen data. Is there any leads or approac

Jason Brownlee October 2, 2018 a

I don't understand, sorry. Training

Your Start in Machine Learning ✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

My3 October 15, 2018 at 10:11 pm #

Hi Jason,

I have some requirement to integrate python code with Java.

I have a ML model which is trained as saved as pickle file, Randomforestclassifier.pkl. I want to load this one time using java and then execute my "prediction" part code which is written python. So my workflow is like:

1. Read Randomforestclassifier.pkl file (one time)
2. Send this model as input to function defined in "python_file.py" which is executed from java for each request
3. python_file.py has prediction code and predictions returned should be captured by java code
Please provide suggestions for this workflow requirement I have used processbuilder in java to execute python_file.py and everything works fine except for model loading as one time activity.

Can you help me with some client server python programming without using rest APIs for one time model loading?

Jason Brownlee October 16, 2018 at 6:37 am #

I recommend treating it like any other engineering project, gather requirements, review options, minimize risk.

Rahul October 18, 2018 at 6:10 pm #   REPLY ↩

Hi Jason,My3,

I have a similar requirement to integrate java with python as my model is in python and in my project we are using java.

Could you please help here.

Jason Brownlee October 19, 2018 at 6:01 am #   REPLY ↩

Thanks for the suggestion.

Theekshana October 30, 2018 at 12:35 am #

Hi Jason,

I have trained my model and evaluated the accuracy u

After evaluating the model, should I train my model wit
model for new future data. (assuming the new model p
from cross-validation)

Thank you for your tutorials and instant replies to questions. 🙂

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee October 30, 2018 at 6:03 am #   REPLY ↩

Yes, see this post:

https://machinelearningmastery.com/train-final-machine-learning-model/

Gagan December 11, 2018 at 5:56 pm #   REPLY ↩

Jason, thanks so much for value add.

Jason Brownlee December 12, 2018 at 5:50 am #   REPLY ↩

You're welcome.

Roger January 1, 2019 at 5:55 am #   REPLY ↩

That helped me a lot. Thank you

Jason Brownlee January 1, 2019 at 6:29 am #   REPLY ↩

I'm happy to hear that.

Kiril Kirov January 4, 2019 at 3:19 am #

How would you go about saving and loading
created using FunctionTransformer?

Your Start in Machine Learning

×

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee January 4, 2019 at 6:33 am

Perhaps pickle?

Shubham January 4, 2019 at 8:37 am #

Hey Jason,

I have a very basic question, let's say I have one model trained on 2017-2018, and then after 6 months I feel
to retrain it on new data. What does retraining actually means here, do I need to have target for my new
data and needs to trained from scratch for new time period, I obviously don't have the target and then how
model will learn from new data.

Jason Brownlee January 4, 2019 at 11:01 am #   REPLY ↩

You have many options, e.g. develop a new model, update the old model, some mixture of the
two with an ensemble.

Rajesh Mahajan January 18, 2019 at 7:44 am #   REPLY ↩

Hi Jason,

I am new to this.. So pardon, if I am asking something incorrect…

I have two stages. Build model and predict.

For Build model:
I am using vectorizer.fit_transform(data) and building the logistic model. My data is a bunch of comments and the target is a set of categories. In order for me to use that model for predicting categories for new comments, I am using the vector created earlier during building of model to predict

So, when I do the save model joblib.dump(log_model, "model.sav")

Foe Predict:
When I try to re-run the model (saved) at a later point of time, I don't have the original vectorizer anymore with the original data set

log_model = joblib.load("model.sav")
inputfeatures_nd = vectorizer.transform(newComment)
pred = log_model.predict(inputfeatures_nd)

I get this error – sklearn.exceptions.NotFittedError: Co

What do you suggest I should do ? Should I be serializ

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee January 18, 2019 at 10:15 a

You must use the same vectorizer that wa
your model.

Rajesh Mahajan January 18, 2019 at 1:17 pm #   REPLY ↩

Thanks Jason! Yes it worked after I save and reload.

Jason Brownlee January 19, 2019 at 5:32 am #   REPLY ↩

I'm happy to hear that.

Ahmed Sahlol February 13, 2019 at 8:23 pm #   REPLY ↩

Thanks Jason for your intersting subjects.
I have this error when saving a VGG16 model after training and testing on my own dataset (can't pickle _thread.RLock objects) when applying the two methods. I also read somewhere that Keras models are not Pickable. So, do you think that those methods are applicable in my case?

Jason Brownlee February 14, 2019 at 8:43 am #

Yes, don't use pickle for Keras models.

Use the Keras save API:

https://machinelearningmastery.com/save-load-keras-deep-learning-models/

shashank February 15, 2019 at 12:24 am #

Awesome post! Keep doing the good work bro!

Jason Brownlee February 15, 2019 at 8:06 am

Thanks.

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

┌───┐
│ Email Address   │
└───┘

START MY EMAIL COURSE

Nick February 19, 2019 at 2:31 am #

Hi Jason, your promised "free" book never ca

🙂

Jason Brownlee February 19, 2019 at 7:27 am #

Sorry to hear that, I can confirm that your email is not in the system, perhaps a typo when you entered it?

Nevertheless, email me directly and I will send you whichever free ebook you are referring to:

https://machinelearningmastery.com/contact/

mayank March 6, 2019 at 7:31 am #

Hi Jason, I have a .sav file where my random forest model has been trained. I have to get back the whole python script for training the model from that .sav file. Is it possible??

Jason Brownlee March 6, 2019 at 8:02 am #

You can load the saved model and start using it.

Rimsha March 29, 2019 at 5:05 am #

Hi Jason, I have trained a model of Naved Baise for sentiment analysis through a trained dataset file of .csv and now I want to use that model for check sentiments of the sentences which are also saved in another .csv file, how could I use?

Jason Brownlee March 29, 2019 at 8:44 am #

Save the model, then load it in a new exa

Nisha March 29, 2019 at 11:23 pm #

Hi Jason,

I have a Class Layer defined to do some functions in K
try to unpickle it, I see an error saying- unknown layer
How should be pickle the model in this case?

Your Start in Machine Learning

✕

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee March 30, 2019 at 6:28 am #

I don't recommend using pickle for Keras models, instead Keras has it's own save model functions:

https://machinelearningmastery.com/save-load-keras-deep-learning-models/

Sevval May 8, 2019 at 5:56 pm #

Can i use my previously saved model for prediction ?

Jason Brownlee May 9, 2019 at 6:38 am #

Yes, load it and call model.predict().

See this post:

https://machinelearningmastery.com/how-to-make-classification-and-regression-predictions-for-deep-learning-models-in-keras/

Sara May 13, 2019 at 1:37 pm #

When in this article you say:

"You might manually output the parameters of your learned model so that you can use them directly in scikit-learn",

I see that we can manually get the tuned hyperparameters, or for example in svm, we can get weight coefficients (coef_),

BUT, is it possible to get svm hyperplane parameters, w and b (y=wx+b) for future predictions?

I believe that pickle saves and load the learned model including w and b but is there any way we can manually output w and b and see what they are in scikit learn?

Many thanks

Jason Brownlee May 13, 2019 at 2:33 pm #

I believe they will be accessible as attribu

You might need to take a closer look at the API or how they are specifically used by sklearn to make

Thank god for open source though, it's all there fo

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Samuel May 18, 2019 at 2:55 am #

Hi, thanks for this helpful article. I'm a beginner and I need to do documents classification. In my model I use :

training_pipeline_data = [
('vectorizer', _create_vectorizer(lang)),
('densifier', _create_densifier()),
('scaler', _create_scaler()),
('classifier', _create_classifier())
]
training_pipeline = ibpip.Pipeline(training_pipeline_data)
training_pipeline.fit(features, labels)

with

def _create_vectorizer(language):
stop_words = safe_get_stop_words(language) if language != 'en' else 'english'
return TfidfVectorizer(sublinear_tf=True, min_df=7, norm='l2', ngram_range=(1, 2),
encoding='latin-1', max_features=500, analyzer='word',
stop_words=stop_words)

def _create_densifier():
return FunctionTransformer(lambda x: x.todense(), accept_sparse=True, validate=False)

```
def _create_scaler():
return StandardScaler()

def _create_classifier():
return GradientBoostingClassifier(n_estimators=160, max_depth=8, random_state=0)
```

Do I have to save in the pickel file the whole pipeline or just the classifier ?

When I save the whole pipeline, the size of the pickel file increases with the amount of training data, but I thouht it shouldn't impact the model size (only the parameters of the model should impact the size of this one)

Jason Brownlee May 18, 2019 at 7:40 am #

Pickle all of it.

Yes, you are saving the mapping of words to numb   encode new samples in the future.

Krtin Ahuja June 17, 2019 at 9:54 pm #

Hi Jason,
How can I load the model to predict further?

Your Start in Machine Learning

×

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee June 18, 2019 at 6:39 am #   REPLY ↩

I show how to load the model in the above tutorial.

What problem are you having exactly?

Raphael June 21, 2019 at 8:00 pm #   REPLY ↩

Hi, big fan of your tutorials.
What are you thought about ONNX (https://onnx.ai/)
Do you think about making a tutorial to explain how it works and how to use it ?

Jason Brownlee June 22, 2019 at 6:38 am #   REPLY ↩

What is ONNX? Perhaps you can summarize it for me?

Constantine June 24, 2019 at 1:43 am #

Hi, thanks for the very useful post, as always!

I wanted to ask you, does this procedure work for saving Grid – Searched models as well? Because when I try to save the grid-search.best_estimator_ it does not give me the results I expect it to (ie the same score on the sample data I use) and the solutions I have found don't work either. Any tips on how to do that?

Many thanks!

Jason Brownlee June 24, 2019 at 6:35 am #

Typically we discard grid search models a   fit a new final model.

Constantine June 24, 2019 at 5:26 pm #

Could you please point me to a source which
search) the following and it does not give me the expe

grid_elastic = GridSearchCV(elastic, param_grid_elasti
cv=tscv.split(X),scoring='neg_mean_absolute_error', v
grid_elastic.fit(X,y)
print(grid_elastic.score(X,y))
filename = 'finalized_model_grid.sav'
joblib.dump(grid_elastic.best_params_, filename,compress=1)
loaded_params_grid = joblib.load(filename)
elastic = ElNet().set_params(**loaded_params_grid)
elastic.fit(X,y)
result = elastic.score(X, y)
print(result)

I grid search an example model, fit it, calculate an example metric for comparisons, and then attempt to save the parameters and use them to instantiate the best estimator later, to avoid having to redo the exhaustive search. It does not give me the same score though. What is wrong? I 've tried just saving the best_estimator_ but it gives me the same wrong result.

Jason Brownlee June 25, 2019 at 6:14 am #

Great question, and this is very common.

Machine learning algorithms are stochastic and we must average their performance over multiple runs.

Your Start in Machine Learning

×

You can master applied Machine Learning **without math or fancy degrees**.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

You can learn more here:
https://machinelearningmastery.com/faq/single-faq/why-do-i-get-different-results-each-time-i-run-the-code

ishrat July 3, 2019 at 9:59 pm #

firstly, thank you for sharing such amazing information always.
this is my code:

```
import time
import numpy as np
import pandas as pd
from nltk import word_tokenize
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.preprocessing import LabelEncoder
from collections import defaultdict
from nltk.corpus import wordnet as wn
from sklearn.feature_extraction.text import TfidfVector
from sklearn import model_selection, svm
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier
import pickle

start_time = time.time()
np.random.seed(500)

#getting only the required columns and rows
dataset = pd.read_csv("records.csv", sep="\t")
dataset_new = dataset.iloc[:, [4, 5, 6, 8, 9]]

df = dataset_new.dropna(subset=['Debit'])
df_required = df.iloc[:, [0, 2]]
df_required = df_required[df_required['Description'] != 'OPENING BALANCE']
df_less = df_required.iloc[:,:]
df_less = df_less.reset_index(drop=True)

# dataset cleanup
df_less = df_less.dropna(subset=['Description'])
df_less = df_less.dropna(subset=['First Level Category'])
df_less['description'] = " "

for index, row in df_less.iterrows():
row['description'] = row['Description'].replace("-", " ")
row['description'] = row['description'].replace("/", " ")
row['description'] = row['description'].replace("_", " ")
```

Your Start in Machine Learning

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

```
row['description'] = row['description'].replace(":", " ")
row['description'] = row['description'].replace(".", " ")

dataset_time = time.time()
print ("Time taken to create dataset : ", dataset_time – start_time)

df_less['description'] = [entry.lower() for entry in df_less['description']]
df_less['description'] = [word_tokenize(entry) for entry in df_less['description']]
df_less = df_less.reset_index(drop=True)

tokenize_time = time.time()
print ("Time taken to tokenize dataset : ", tokenize_time – dataset_time)

for index, entry in enumerate(df_less['description']):
Final_words = []
for word in entry:
if word.isalpha():
Final_words.append(word)
df_less.loc[index, 'desc_final'] = str(Final_words)

df_others = df_less[df_less['desc_final'] == '[]']
df_less_final = pd.DataFrame()
df_less_final = df_less[df_less['desc_final'] != '[]']
df_less_final = df_less_final.reset_index(drop=True)
data_cleanup_time = time.time()
print ("Time taken for data cleanup", data_cleanup_tim

Train_X, Test_X, Train_Y, Test_Y = model_selection.trai
df_less_final['First Level Category'], test_size=0.33,
random_state=10,shuffle=True)

Tfidf_vect = TfidfVectorizer(max_features=106481)
Tfidf_vect.fit(df_less['desc_final'])
Train_X_Tfidf = Tfidf_vect.transform(Train_X)
Test_X_Tfidf = Tfidf_vect.transform(Test_X)

clf = RandomForestClassifier(n_jobs=8, random_state=10)

clf.fit(Train_X_Tfidf, Train_Y)
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=None, max_features='auto', max_leaf_nodes=None,
min_impurity_split=1e-07, min_samples_leaf=20,
min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=100, n_jobs=8, oob_score=False, random_state=10,
verbose=0, warm_start=False)
preds = clf.predict(Test_X_Tfidf)
print("Random forest Accuracy Score -> ", accuracy_score(preds, Test_Y) * 100)
preds.tofile("foo.csv", sep = '\n')
dff = pd.DataFrame()
col_name = ['category']
dff = pd.read_csv("foo.csv", names = col_name, sep="\n")
```

sir this a model that i have prepared now i want to dump it using pickle but i am not able to understand how can i do this…since everytime i want to predict new records i want to preprocess my one of my rows as i am doing above, and also used vectorizer..and then predict the results..can you please help me with the solution.

thank you

Jason Brownlee July 4, 2019 at 7:47 am #   REPLY ↩

This is a common question that I answer here:

https://machinelearningmastery.com/faq/single-faq/can-you-read-review-or-debug-my-code

teimoor July 17, 2019 at 4:14 am #

hi how can i learn python fast for the purpose
can you notify me on gmail please

Your Start in Machine Learning   ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

Jason Brownlee July 17, 2019 at 8:30 am #

Right here:
http://machinelearningmastery.com/crash-course-

Raghad July 22, 2019 at 5:59 pm #   REPLY ↩

Thank you so much for all your effort, but I am a beginner in ML and Python and I have a basic conceptual question:
I used a CSV file to train, test and fit my random forest model then I saved the model in a pickle file. Now my partner wants to use the model for prediction on new unseen data(entered by user) so my question is should I send her only the model I saved in a pickle file or also the data I used to train and fit the model? I mean would the pkl model work even if the CSV file containing the data used to fit the model is not in the same folder or host? I hope my question is clear. Thank you again very much!!

Jason Brownlee July 23, 2019 at 7:57 am #   REPLY ↩

Just the model is required.

Leave a Reply

Name (required)

Email (will not be published) (requ

Website

SUBMIT COMMENT

Welcome to Machine Learning Mastery!

Hi, I'm **Jason Brownlee**, PhD.
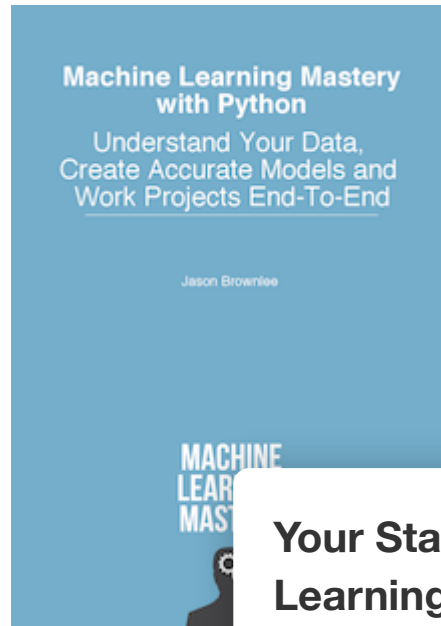I write tutorials to help developers (*like you*) get results with machine learning.

Read More

More Tutorials?

Start Here

(over 780 tutorials!)

Machine Learning with Python
Develop predictive models with scikit-learn.

Click to Get Started Now!

Machine Learning Mastery
with Python
Understand Your Data,
Create Accurate Models and
Work Projects End-To-End

Jason Brownlee

MACHINE
LEAR
MAST

Join Nev

Subscri

(join over 150,

Your Start in Machine Learning ✕

You can master applied Machine Learning
without math or fancy degrees.
Find out how in this *free* and *practical* course.

Email Address

START MY EMAIL COURSE

RSS | Twitter | Facebook | LinkedIn

Privacy | Disclaimer | Terms | Contact | Sitemap | Search