☰   **Navigation**

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

Search...	🔍

Feature Selection For Machine Learning in Python

by **Jason Brownlee** on May 20, 2016 in **Python Machine Learning**

Tweet   Share   Share

The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

Irrelevant or partially relevant features can negatively impact model performance.

In this post you will discover automatic feature selection techniques that you can use to prepare your machine learning data in python with scikit-learn.

Let's get started.

- **Update Dec/2016**: Fixed a typo in the RFE section regarding the chosen variables. Thanks Anderson.
- **Update Mar/2018**: Added alternate link to download the dataset as the original appears to have been taken down.

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

Feature Selection For Machine Learning in Python
Photo by Baptiste Lafontaine, some rights reserved.

Feature Selection

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

Three benefits of performing feature selection before modeling your data are:

- **Reduces Overfitting**: Less redundant data mean[s] noise.
- **Improves Accuracy**: Less misleading data mean[s]
- **Reduces Training Time**: Less data means that a[ ]

You can learn more about feature selection with sciki[t]

Need help with Machin[e]

Take my free 2-week email course and discove[r]

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

Feature Selection for Machine Learning

This section lists 4 feature selection recipes for machine learning in Python

This post contains recipes for feature selection methods.

Each recipe was designed to be complete and standalone so that you can copy-and-paste it directly into you project and use it immediately.

Recipes uses the Pima Indians onset of diabetes dataset to demonstrate the feature selection method (update: download from here). This is a binary classification problem where all of the attributes are numeric.

1. Univariate Selection

Statistical tests can be used to select those features that have the strongest relationship with the output variable.

The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.

The example below uses the chi squared (chi^2) statistical test for non-negative features to select 4 of the best features from the Pima Indians onset of diabetes dataset.

```
1   # Feature Extraction with Univariate Statistical Tests (Chi-squared for classification)
2   import pandas
3   import numpy
4   from sklearn.feature_selection import SelectKBest
5   from sklearn.feature_selection import chi2
6   # load data
7   url = "https://raw.githubusercontent.com/jbrow                    ta.c
8   names = ['preg', 'plas', 'pres', 'skin', 'tes
9   dataframe = pandas.read_csv(url, names=names)
10  array = dataframe.values
11  X = array[:,0:8]
12  Y = array[:,8]
13  # feature extraction
14  test = SelectKBest(score_func=chi2, k=4)
15  fit = test.fit(X, Y)
16  # summarize scores
17  numpy.set_printoptions(precision=3)
18  print(fit.scores_)
19  features = fit.transform(X)
20  # summarize selected features
21  print(features[0:5,:])
```

You can see the scores for each attribute and the 4 attributes chosen (those with the highest scores): *plas, test, mass* and *age*.

```
1  [  111.52    1411.887     17.605      53.108  2175.565    127.669       5.393
2      181.304]
3  [[ 148.       0.     33.6    50. ]
4   [  85.       0.     26.6    31. ]
5   [ 183.       0.     23.3    32. ]
6   [  89.      94.     28.1    21. ]
7   [ 137.     168.     43.1    33. ]]
```

2. Recursive Feature Elimination

The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain.

It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

You can learn more about the RFE class in the scikit-learn documentation.

The example below uses RFE with the logistic regression algorithm to select the top 3 features. The choice of algorithm does not matter too much as long as it is skillful and consistent.

```
1  # Feature Extraction with RFE
2  from pandas import read_csv
3  from sklearn.feature_selection import RFE
4  from sklearn.linear_model import LogisticRegression
5  # load data
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 = read_csv(url, names=names)
9  array = dataframe.values
10 X = array[:,0:8]
11 Y = array[:,8]
12 # feature extraction
13 model = LogisticRegression()
14 rfe = RFE(model, 3)
15 fit = rfe.fit(X, Y)
16 print("Num Features: %d") % fit.n_features_
17 print("Selected Features: %s") % fit.support_
18 print("Feature Ranking: %s") % fit.ranking_
```

You can see that RFE chose the the top 3 features as

These are marked True in the *support_* array and mark

```
1  Num Features: 3
2  Selected Features: [ True False False False Fal
3  Feature Ranking: [1 2 3 5 6 1 1 4]
```

3. Principal Component Analysis

Principal Component Analysis (or PCA) uses linear alg
form.

Generally this is called a data reduction technique. A property of PCA is that you can choose the number of dimensions or principal component in the transformed result.

In the example below, we use PCA and select 3 principal components.

Learn more about the PCA class in scikit-learn by reviewing the PCA API. Dive deeper into the math behind PCA on the Principal Component Analysis Wikipedia article.

```
1  # Feature Extraction with PCA
2  import numpy
3  from pandas import read_csv
4  from sklearn.decomposition import PCA
5  # load data
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 = read_csv(url, names=names)
9  array = dataframe.values
10 X = array[:,0:8]
11 Y = array[:,8]
12 # feature extraction
13 pca = PCA(n_components=3)
14 fit = pca.fit(X)
15 # summarize components
16 print("Explained Variance: %s") % fit.explained_variance_ratio_
17 print(fit.components_)
```

You can see that the transformed dataset (3 principal components) bare little resemblance to the source data.

```
1  Explained Variance: [ 0.88854663  0.06159078  0.02579012]
2  [[ -2.02176587e-03   9.78115765e-02   1.60930503e-02   6.07566861e-02
3     9.93110844e-01   1.40108085e-02   5.37167919e-04  -3.56474430e-03]
4   [ 2.26488861e-02   9.72210040e-01   1.41909330e-01  -5.78614699e-02
5    -9.46266913e-02   4.69729766e-02   8.16804621e-04   1.40168181e-01]
6   [ -2.24649003e-02   1.43428710e-01  -9.22467192e-01  -3.07013055e-01
7     2.09773019e-02  -1.32444542e-01  -6.39983017e-04  -1.25454310e-01]]
```

4. Feature Importance

Bagged decision trees like Random Forest and Extra Trees can be used to estimate the importance of features.

In the example below we construct a ExtraTreesClass[...]es dataset. You can learn more about the ExtraTreesClas[...]

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  # Feature Importance with Extra Trees Classifi
2  from pandas import read_csv
3  from sklearn.ensemble import ExtraTreesClassif
4  # load data
5  url = "https://raw.githubusercontent.com/jbrow
6  names = ['preg', 'plas', 'pres', 'skin', 'test
7  dataframe = read_csv(url, names=names)
8  array = dataframe.values
9  X = array[:,0:8]
10 Y = array[:,8]
11 # feature extraction
12 model = ExtraTreesClassifier()
13 model.fit(X, Y)
```

```
14 print(model.feature_importances_)
```

You can see that we are given an importance score for each attribute where the larger score the more important the attribute. The scores suggest at the importance of *plas*, *age* and *mass*.

```
1 [ 0.11070069  0.2213717   0.08824115  0.08068703  0.07281761  0.14548537 0.12654214  0.15415431]
```

Summary

In this post you discovered feature selection for preparing machine learning data in Python with scikit-learn.

You learned about 4 different automatic feature selection techniques:

- Univariate Selection.
- Recursive Feature Elimination.
- Principle Component Analysis.
- Feature Importance.

If you are looking for more information on feature selection, see these related posts:

- Feature Selection with the Caret R Package
- Feature Selection to Improve Accuracy and Decrease Training Time
- An Introduction to Feature Selection
- Feature Selection in Python with Scikit-Learn

Do you have any questions about feature selection or this post? Ask your questions in the comment and I will do my best to answer them.

Frustrated With Python Machine Learning?

Develop Your Own Models in Minutes

…with just a few line

Discover how in
Machine Learning M

Covers **self-study tutorials** a
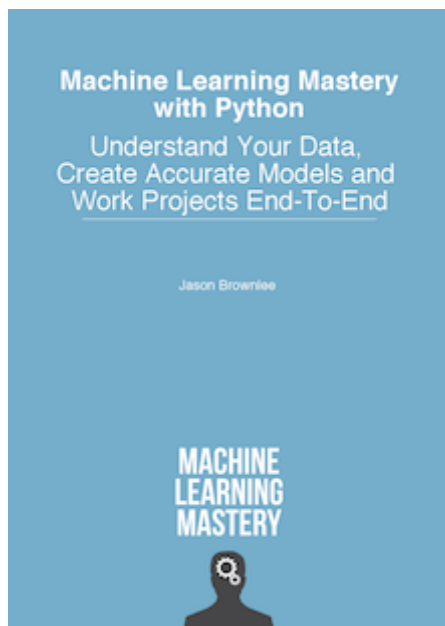*Loading data*, *visualization*, *mod*

Finally Bring Mac
Your Own

Skip the Academ

Click to le

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

**Machine Learning Mastery
with Python**

Understand Your Data,
Create Accurate Models and
Work Projects End-To-End

Jason Brownlee

**MACHINE
LEARNING
MASTERY**

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 →

‹ How To Build Multi-Layer Perceptron Neural Network Models with Keras

Evaluate the Performance of Machine Learning Algorithms in Python using Resampling ›

248 Responses to *Feature Selection For Machine Learning in Python*

Juliet September 16, 2016 at 8:57 pm #

Hi Jason! Thanks for this – really useful post!
looking at your Univariate Analysis, the features you ha
the highest values in the printed score summary. Is tha
Thanks again for a great access-point into feature sele

**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 17, 2016 at 9:29

Hi Juliet, it might just be coincidence. If you uncover something different, please let me know.

Ansh October 11, 2016 at 12:16 pm #   REPLY ↰

For the Recursive Feature Elimination, are the features of high importance (preg,mass,pedi)? The ranking array has value 1 for them them.

Jason Brownlee October 12, 2016 at 9:11 am #   REPLY ↰

Hi Ansh, I believe the features with the 1 are preg, pedi and age as mentioned in the post. These are the first ranked features.

Ansh October 12, 2016 at 12:29 pm #   REPLY ↰

Thanks for the reply Jason. I seem to have made a mistake, my bad. Great post 🙂

Jason Brownlee October 13, 2016 at 8:33 am #   REPLY ↰

No problem Ansh.

Anderson Neves December 15, 2016 at 6:52 am #   REPLY ↰

Hi all,

I agree with Ansh. There are 8 features and the indexes with True and 1 match with preg, mass and pedi.

['preg', 'plas', 'pres', 'skin', 'test', 'ma
[True, False, False, False, False, True,
[1, 2, 3, 5, 6, 1, 1, 4]

Jason, could you explain better how y
features?

Thank you for the post, it was very use

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 15

Hi Anderson, they have a "true" in their column index and are all ranked "1" at their respective column index.

Does that help?

Anderson Neves December 16, 2016 at 12:00 am #

Hi Jason,

That is exactly what I mean. I believe that the best features would be preg, pedi and age in the scenario below

Features:
['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age']

RFE result:
[True, False, False, False, False, False, True, True]
[1, 2, 3, 5, 6, 4, 1, 1]

However, the result was

Features:
['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age']

RFE result:
[True, False, False, False, False, True, True, False]
[1, 2, 3, 5, 6, 1, 1, 4]

Did you consider the target column 'class' by mistake?

Thank you for the quick reply,
Anderson Neves

Jason Brownlee December 16, 2016 at 5:48 am #

Hi Anderson,

I see, you're saying you have a differer

The code is correct and does not inclu

Re-running now I see the same result:

```
1  Num Features: 3
2  Selected Features: [ True Fals
3  Feature Ranking: [1 2 3 5 6 1
```

Perhaps I don't understand the proble

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

Anderson Neves December 17, 2016 at 12:22 am #

Hi Jason,

Your code is correct and my result is the same as yours. My point is that the best features found with RFE are preg, mass and pedi. So, I suggest you fix the text "You can see that RFE chose the the top 3 features as preg, pedi and age.". If you add the code below at the end of your code you will see what I mean.

```
# find best features
best_features = []
i = 0
for is_best_feature in fit.support_:
if is_best_feature:
best_features.append(names[i])
i += 1
print '\nSelected features:'
print best_features
```

Sorry if I am bothering somehow,
Thanks again,
Anderson Neves

Jason Brownlee December 17, 2016 at 11:18 am #

Got it Anderson.
Thanks for being patient with me and helping to make this post more useful. I really appreciate it!

I've fixed up the example above.

REPLY

Narasimman October 14, 2016 at 9:18 pm #

from the rfe, how do I form a new dataframe f

Jason Brownlee October 15, 2016 at 10:22 a

Great question Narasimman.

From memory, you can use numpy.concatinate() tc

http://docs.scipy.org/doc/numpy/reference/genera

×

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

Iain Dinwoodie November 1, 2016 at 12:52 am #   REPLY ↩

Thanks for useful tutorial.

Narasimman – 'from the rfe, how do I form a new dataframe for the features which has true value?'

You can just apply rfe directly to the dataframe then select based on columns:

…
df = read_csv(url, names=names)
X = df.iloc[:, 0:8]
Y = df.iloc[:, 8]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, 3)
fit = rfe.fit(X, Y)
print("Num Features: {}".format(fit.n_features_))
print("Selected Features: {}".format(fit.support_))
print("Feature Ranking: {}".format(fit.ranking_))

X = X[X.columns[fit.support_]]

MLBeginner October 25, 2016 at 1:07 am #   REPLY ↩

Hi Jason,

Really appreciate your post! Really great! I have a quick question for the PCA method. How to get the column header for the selected 3 principal components? It is just simple column no. there, but hard to know which attributes finally are.

Thanks,

Jason Brownlee October 25, 2016 at 8:29 am #   REPLY ↩

Thanks MLBeginner, I'm glad you found it

There is no column header, they are "new" feature:

sadiq October 25, 2016 at 1:51 am #

hi, Jason! please I want to ask you if i can use

python

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 25, 2016 at 8:29 am #

Sure, try it and see how the results compare (as in the models trained on selected features) to other feature selection methods.

Vignesh Sureshbabu Kishore November 15, 2016 at 5:07 pm #

Hey Jason, can the univariate test of Chi2 feature selection be applied to both continuous and categorical data.

Jason Brownlee November 16, 2016 at 9:25 am #

Hi Vignesh, I believe just continuous data. But I may be wrong – try and see.

Vignesh Sureshbabu Kishore November 16, 2016 at 1:07 pm #

Hey Jason, Thanks for the reply. In the univariate selection to perform the chi-square test you are fetching the array from df.values. In that case, each element of the array will be each row in the data frame.

To perform feature selection, we should have ideally fetched the values from each column of the dataframe to check the independence of each feature with the class variable. Is it a inbuilt functionality of the sklearn.preprocessing beacuse of which you fetch the values as each row.

Please suggest me on this.

Jason Brownlee November 17, 2016 at 9:49 am #

I'm not sure I follow Vignesh. Gen the tests.

Vineet December 2, 2016 at 5:11 am #

Hi Jason,

I am trying to do image classification using cpu machi means 200000 features. Pls suggest how do I reduce

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 2, 2016 at 8:19 am #

Consider working with a sample of the dataset.

Consider using the feature selection methods in this post.

Consider projection methods like PCA, sammons mapping, etc.

I hope that helps as a start.

---

tvmanikandan December 15, 2016 at 5:49 pm #

Jason,
when you use "SelectKBest" , can you please explain how you get the below scores?

[111.52 1411.887 17.605 53.108 2175.565 127.669 5.393
181.304]

-Mani

Jason Brownlee December 16, 2016 at 5:40 am #

I use a chi squared test, you can learn more about it here:

http://scikit-
learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2

---

tvmanikandan December 16, 2016 at 5:29 pm #

Jason,
I understand you used chi square. But if want to get these scores manually , how can i do it? Please
explain.

-Mani

Jason Brownlee December 17, 201

Good question, I don't have an ex

**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

tvmanikandan December 16, 2016 at 2:48 am #

jason,

Please explain how the below scores are achieved using chi2.

[111.52 1411.887 17.605 53.108 2175.565 127.669 5.393
181.304]

-Mani

Natheer Alabsi December 28, 2016 at 8:35 pm #   REPLY ↰

Jason, how can we get feature names from their rankings?

Jason Brownlee December 29, 2016 at 7:15 am #   REPLY ↰

Hi Natheer,

Map the feature rank to the index of the column name from the header row on the DataFrame or
whathaveyou.

Jason January 9, 2017 at 2:40 am #   REPLY ↰

Hi Jason,

Thank you for this nice blog

I have a regression problem and I need to convert a bunch of categorical variables into dummy data, which
will generate over 200 new columns. Should I do the feature selection before this step or after this step?
Thanks

Jason Brownlee January 9, 2017 at 7:52 am #   REPLY ↰

Try and see.

That is a lot of new binary variables. Your resulting
selection prior might be a good idea, also try after.

Mohit Tiwari February 13, 2017 at 3:37 pm #

Hi Jason,

I am bit stuck in selecting the appropriate feature selec

I have about 900 attributes (columns) in my data and about 60 records. The values are nothing but count of attributes.

Basically, I am taking count of API calls of a portable file.

My data is like this:

File, dangerous, API 1,API 2,API 3,API 4,API 5,API 6…..API 900
ABC, yes, 1,0,2,1,0,0,….
DEF, no,0,1,0,0,1,2
FGH,yes,0,0,0,1,2,3

.

.

.

Till 60

Can u please suggest me a suitable feature selection for my data?

Jason Brownlee February 14, 2017 at 10:03 am #

Hi Mohit,

Consider trying a few different methods, as well as some projection methods and see which "views" of your data result in more accurate predictive models.

Esu February 15, 2017 at 12:01 am #

Hell!

Once I got the reduced version of my data as a result of using PCA, how can I feed to my classifier?

example: the original data is of size 100 row by 5000 columns
if I reduce 200 features I will get 100 by 200 dimension data. right?
then I create arrays of

a=array[:,0:199]
b=array[:,99]
but when I test my classifier its core is 0% in both test
An7y Idea

Jason Brownlee February 15, 2017 at 11:35 a

Sounds like you're on the right, but a zero

Did you accidently include the class output variabl
excluded.

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

Kamal February 20, 2017 at 6:20 pm #   REPLY ↰

Hello sir,

I have a question in my mind

each of these feature selection algo uses some predefined number like 3 in case of PCA.So how we come to know that my data set cantain only 3 or any predefined number of features.it does not automatically select no features its own.

Jason Brownlee February 21, 2017 at 9:33 am #   REPLY ↰

Great question Kamal.

No, you must select the number of features. I would recommend using a sensitivity analysis and try a number of different features and see which results in the best performing model.

Massimo March 9, 2017 at 5:29 am #   REPLY ↰

Hi jason,

I have a question about the RFECV approach.

I'm dealing with a project where I have to use different estimators (regression models). is it correct use RFECV with these models? or is it enough to use only one of them? Once I have selected the best features, could I use them for each regression model?

To better explain:

– I have used RFECV on whole dataset in combination with one of the following regression models [LinearRegression, Ridge, Lasso]

– Then I have compared the r2 and I have chosen the better model, so I have used its features selected in order to do others things.

– pratically, I use the same 'best' features in each regression model.

Sorry for my bad english.

Jason Brownlee March 9, 2017 at 9:58 am #

Good question.

You can embed different models in RFE and see if
of what features to pick.

You can build a model from each set of features ar

You can pick one set of features and build one or r

My advice is to try everything you can think of and see what gives the best results on your validation dataset.

Massimo March 11, 2017 at 2:41 am #

Thank you man. You're great.

Jason Brownlee March 11, 2017 at 8:01 am #

You're welcome.

gevra March 22, 2017 at 1:49 am #

Hi Jason.

Thanks for the post, but I think going with Random Forests straight away will not work if you have correlated features.

Check this paper:
https://academic.oup.com/bioinformatics/article/27/14/1986/194387/Classification-with-correlated-features

I am not sure about the other methods, but feature correlation is an issue that needs to be addressed before assessing feature importance.

Jason Brownlee March 22, 2017 at 8:08 am #

Makes sense, thanks for the note and the reference.

ssh June 20, 2017 at 8:20 pm #

Jason, following this notes, do you ha
input vectors become problematic in the mach
technics which embedded in some algos (like
supply some answer to the correlations issue.
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 June 21, 2017 at 8

Perhaps a correlation above 0.5. Perform a sensitivity analysis with different values, select features and use resulting model skill to help guide you.

ogunleye March 30, 2017 at 4:29 am #

Hello sir,

Thank you for the informative post. My questions are

1) How do you handle NaN in a dataset for feature selection purposes.

2) I am getting an error with RFE(model, 3) It is telling me i supplied 2 arguments instead of 1.

Thank you very much once again.

Jason Brownlee March 30, 2017 at 8:57 am #

Hi, NaN is a mark of missing data.

Here are some ways to handle missing data:

http://machinelearningmastery.com/handle-missing-data-python/

ogunleye March 30, 2017 at 4:33 am #

I solved my problem sir. I named the function RFE in my main but. I would love to hear your response to first question.

Sam April 20, 2017 at 3:49 am #

how to load the nested JSON into the data frame ?

Jason Brownlee April 20, 2017 at 9:32 am #

I don't know off hand, perhaps post to St

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

Federico Carmona April 20, 2017 at 6:10 am #

good afternoon

How to know with pca what are the main components

Jason Brownlee April 20, 2017 at 9:34 am #

PCA will calculate and return the principal components.

Federico Carmona April 20, 2017 at 10:53 am #

Yes but pca does not tell me which are the most relevant varials if mass test etc?

Jason Brownlee April 21, 2017 at 8:27 am #

Not sure I follow you sorry.

You could apply a feature selection or feature importance method to the PCA results if you wanted. It might be overkill though.

Lehyu April 23, 2017 at 6:44 pm #

In RFE we should input a estimator, so before I do feature selection, should I fine tune the model or just use the default parmater settting? Thanks.

Jason Brownlee April 24, 2017 at 5:33 am #

You can, but that is not really required. As long as the estimator is reasonably skillful on the problem, the selected features will be valuable.

Lehyu April 25, 2017 at 12:41 am #

I was suck here for days. Thanks a lo

Lehyu April 25, 2017 at 1:09 am #

stuck…

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 April 25, 2017 at 7:49 am #

I'm glad to hear the advice helped.

I'm here to help if you get stuck again, just post your questions.

Rj May 7, 2017 at 4:38 pm #

Hi Jason,

I was wondering if I could build/train another model (say SVM with RBF kernel) using the features from SVM-RFE (wherein the kernel used is a linear kernel).

Jason Brownlee May 8, 2017 at 7:42 am #

Sure.

Gwen June 5, 2017 at 7:02 pm #

Hi Jason,

First of all thank you for all your posts ! It's very helpful for machine learning beginners like me.

I'm working on a personal project of prediction in 1vs1 sports. My neural network (MLP) have an accuracy of 65% (not awesome but it's a good start). I have 28 features and I think that some affect my predictions. So I applied two algorithms mentionned in your post :
– Recursive Feature Elimination,
– Feature Importance.

But I have some contradictions. For exemple with RFE I determined 20 features to select but the feature the most important in Feature Importance is not selected in RFE. How can we explain that ?

In addition to that in Feature Importance all features are [...] features are not correlated with my ouput ?

Thanks again for your help !

Jason Brownlee June 6, 2017 at 9:30 am #

Hi Gwen,

Different feature selection methods will select diffe[...]

Build a model on each set of features and compar[...]

Consider ensembling the models together to see if performance can be lifted.

A great area to consider to get more features is to use a rating system and use rating as a highly predictive input variable (e.g. chess rating systems can be used directly).

Let me know how you go.

Gwen June 7, 2017 at 1:17 am #   REPLY ↰

Thanks for your answer Jason.

I tried with 20 features selected by Recursive Feature Elimination but my accuracy is about 60%…

In addition to that the Elo Rating system (used in chess) is one of my features. With this feature only my accuracy is ~65%.

Maybe a MLP is not a good idea for my project. I have to think about my NN configuration I only have one hidden layer.

And maybe we cannot have more than 65/70% for a prediction of tennis matches.
(Not enough for a positive ROI !)

Jason Brownlee June 7, 2017 at 7:23 am #   REPLY ↰

Hang in there Gwen.

Try lots of models and lots of config for models.

See what skill other people get on the same or similar problems to get a feel for what is possible.

Brainstorm for days about features and other data you could use.

See this post:

http://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/

RATNA NITIN PATIL July 20, 2017 at 8:16 pm #

Hello Jason,

I am very much impressied by this tutorial. I am just a l
reduced version of my data as a result of using PCA, h
feed the output of PCA to build the classifier?

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 21, 2017 at 9:33 am #

Assign it to a variable or save it to file then use the data like a normal input dataset.

RATNA NITIN PATIL July 20, 2017 at 8:56 pm #

Hi Jason,

I was trying to execute the PCA but, I got the error at this point of the code

print("Explained Variance: %s") % fit.explained_variance_ratio_

It's a type error: unsupported operand type(s) for %: 'non type' and 'float'

Please help me.

Jason Brownlee July 21, 2017 at 9:35 am #

Looks like a Python 3 issue. Move the ")" to the end of the line:

```
1  print("Explained Variance: %s" % fit.explained_variance_ratio_)
```

RATNA NITIN PATIL July 21, 2017 at 2:23 pm #

Thanks Jason. It works.

Jason Brownlee July 22, 2017 at 8:29 am #

Glad to hear it.

Raphael Alencar July 21, 2017 at 9:57 pm #

How to know wich feature selection technique

Jason Brownlee July 22, 2017 at 8:35 am #

Consider using a few, create models for e
performing model.

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

RATNA NITIN PATIL July 22, 2017 at 4:23 pm #

Hello Jason,

I have used the extra tree classifier for the feature selection then output is importance score for each attribute. But then I want to provide these important attributes to the training model to build the classifier. I am not able to provide only these important features as input to build the model.
I would be greatful to you if you help me in this case.

Jason Brownlee July 23, 2017 at 6:20 am #

The importance scores are for you. You can use them to help decide which features to use as inputs to your model.

Haiyang Duan December 11, 2018 at 7:57 pm #

Hi Jason, I truely appreciate your post. But I have a quick question. Why the sum of the importance scores is unequal to 1?

Jason Brownlee December 12, 2018 at 5:51 am #

Because they are not normalized.

Haiyang Duan December 12, 2018 at 11:46 am #

I am sincerely grateful to you. I ran feature importance using SelectFromModel with estimator=LinearSVC. But I got negative feature importance values. I would like to kown what that means.

Jason Brownlee December 12

Scores are often relative. Perh

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

Haiyang Duan December 12, 20

Thank you very much.

RATNA NITIN PATIL July 22, 2017 at 6:33 pm #
REPLY ↩

Hi Jason,

Basically i want to provide feature reduction output to Naive Bays. I f you could provide sample code will be better.

Thanks for providing this wonderful tutorial.

Jason Brownlee July 23, 2017 at 6:21 am #
REPLY ↩

You can use feature selection or feature importance to "suggest" which features to use, then develop a model with those features.

RATNA NITIN PATIL July 23, 2017 at 6:44 pm #
REPLY ↩

Thanks Jason,

But after knowing the important features, I am not able to build a model from them. I don't know how to giveonly those featueslimportant) as input to the model. I mean to say X_train parameter will have all the features as input.

Thanks in advance….

Jason Brownlee July 24, 2017 at 6:53 am #
REPLY ↩

A feature selection method will tell you which features you could use. Use your favorite programming language to make a new data file with just those columns.

RATNA NITIN PATIL July 24, 2017 at 5:

thanks a lot Jason. You are doing a g

Jason Brownlee July 25, 2017 at 9:

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

RATNA NITIN PATIL July 24, 2017 at 6:11 pm #

I have my own dataset on the Desktop, not the standard dataset that all machine learning have in their depositories (e.g. iris , diabetes).

I have a simple csv file and I want to load it so that I can use scikit-learn properly.

I need a very simple and easy way to do so.

Waiting for the reply.

Jason Brownlee July 25, 2017 at 9:37 am #

Try this tutorial:
http://machinelearningmastery.com/load-machine-learning-data-python/

mllearn July 29, 2017 at 6:04 am #

Thanks for this post, it's very helpful,

What would make me choose one technique and not the others?
The results of each of these techniques correlates with the result of others?, I mean, makes sense to use more than one to verify the feature selection?.

thanks!

Jason Brownlee July 29, 2017 at 8:12 am #

Choose a technique based on the results of a model trained on the selected features.

In predictive modeling we are concerned with increasing the skill of predictions and decreasing model complexity.

mllearn July 30, 2017 at 5:04 pm #

Sounds that I'd need to cross-validate
depends on the data but I'm trying to figure ou

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 31, 2017 at 8:

Applied machine learning is empirical. You cannot pick the "best" methods analytically.

---

steve August 17, 2017 at 3:15 pm #

Hi Jason,

In your examples, you write:

array = dataframe.values
X = array[:,0:8]
Y = array[:,8]

In my dataset, there are 45 features. When i write like this:

X = array[:,0:44]
Y = array[:,44]

I get some errors:

Y = array[:,44]
IndexError: index 45 is out of bounds for axis 1 with size 0

If you help me, i ll be grateful!
Thanks in advance.

Jason Brownlee August 17, 2017 at 4:55 pm #

Confirm that you have loaded your data correctly, print the shape and some rows.

Aneeshaa S C August 20, 2017 at 11:22 pm #

1.. What kind of predictors can be used with Lasso?
2. If categorical predictors can be used, should they be
values re-coded to be 1/0
3. Can categorical variables such as location (U(urban)

Jason Brownlee August 21, 2017 at 6:07 am

Regression, e.g. predicting a real value.

Categorical inputs must be encoded as integers 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

panteha August 29, 2017 at 1:36 am #

Hi Jason

I am new to ML and am doing a project in Python, at some point it is to recognize correlated features , I wonder what will be the next step? what to do with correlated features? should we change them to something new? a combination maybe? how does it affect our modeling and prediction? appreciated if you direct me into some resources to study and find it out.

best

Jason Brownlee August 29, 2017 at 5:09 pm #

It is common to identify and remove the correlated input variables.

Try it and see if it lifts skill on your model.

Silvio Abela September 26, 2017 at 6:48 am #

Hello Dr Brownlee,

Thank you for these incredible tutorials.

I am trying to classify some text data collected from online comments and would like to know if there is any way in which the constants in the various algorithms can be determined automatically. For example, in SelectKBest, k=3, in RFE you chose 3, in PCA, 3 again whilst in Feature Importance it is left open for selection that would still need some threshold value.

Is there a way like a rule of thumb or an algorithm to automatically decide the "best of the best"? Say, I use n-grams; if I use trigrams on a 1000 instance data set, the number of features explodes. How can I set SelectKBest to an "x" number automatically according to the best? Thank you.

Jason Brownlee September 26, 2017 at 2:59

No, hyperparameters cannot be set analy best configuration for your specific problem.

You can use heuristics or copy values, but really th test harness.

Abby October 6, 2017 at 3:43 pm #

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

It was an impressive tutorial, quite easy to understand. I am looking for feature subset selection using gaussian mixture clustering model in python. Can you help me out?

Jason Brownlee October 7, 2017 at 5:48 am #   REPLY ↩

Sorry, I don't have material on mixture models or clustering. I cannot help.

Manjunat October 6, 2017 at 8:31 pm #   REPLY ↩

Hi jason

I've tried all feature selection techniques which one is opt for training the data for the predictive modelling …?

Jason Brownlee October 7, 2017 at 5:54 am #   REPLY ↩

Try many for your dataset and see which subset of features results in the most skillful model.

Nerea October 16, 2017 at 7:13 pm #   REPLY ↩

Hello Jason,

I am a biochemistry student in Spain and I am on a project about predictive biomarkers in cancer. The bioinformatic method I am using is very simple but we are trying to predict metastasis with some protein data. In our research, we want to determine the best biomarker and the worst, but also the synergic effect that would have the use of two biomarkers. That is my problem: I don't know how to calculate which are the two best predictors.

This is what I have done for the best and worst predictors:

analisis=['il10meta']
X = data[analisis].values

#response variable
response='evol'
y = data[response].values

use train/test split with different random_state values
X_train, X_test, y_train, y_test = train_test_split(X, y, ra

from sklearn.neighbors import KNeighborsClassifier

#creating the classfier
knn = KNeighborsClassifier(n_neighbors=1)

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

#fitting the classifier
knn.fit(X_train, y_train)

#predicting response variables corresponding to test data
y_pred = knn.predict(X_test)

#calculating classification accuracy
print(metrics.accuracy_score(y_test, y_pred))

I have calculate the accuracy. But when I try to do the same for both biomarkers I get the same result in all the combinations of my 6 biomarkers.

Could you help me? Any tip?

THANK YOU

Jason Brownlee October 17, 2017 at 5:41 am #   REPLY ↩

Generally, I would recommend following this process to get the best model for your predictive modeling problem:

https://machinelearningmastery.com/start-here/#process

Generally, you must test many different models and many different framings of the problem to see what works best.

gen October 17, 2017 at 6:35 pm #   REPLY ↩

Hello Jason,

Many thanks for your post. I have also read your introduction article about feature selection. Which method is Feature Importance categorized under? i.e wrapper or embedded ?

Thanks

Jason Brownlee October 18, 2017 at 5:32 am

Neither, it is a different thing yet again.

You could use the importance scores as a filter.

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

Numan Yilmaz October 26, 2017 at 1:46 pm #

Great post! Thank you, Jason. My question is
for all algorithms. What if I have categorical data? How

model if there are categorical features? Is there a method/way to calculate it before one-hot encoding(get_dummies) or how to calculate after one-hot encoding if the model is not tree-based?

Jason Brownlee October 26, 2017 at 4:18 pm #   REPLY ↩

Good question, I cannot think of feature selection methods specific to categorical data off hand, they may be out there. Some homework would be required (e.g. google scholar search).

rohit November 13, 2017 at 9:11 pm #   REPLY ↩

hello Jason,
Should I do Feature Selection on my validation dataset also? Or just do feature selection on my training set alone and then do the validation using the validation set?

Jason Brownlee November 14, 2017 at 10:10 am #   REPLY ↩

Use the train dataset to choose features. Then, only choose those features on test/validation and any other dataset used by the model.

Maryam November 16, 2017 at 3:45 pm #   REPLY ↩

hello jason
i am doing simple classification but there is coming an issue
ValueError Traceback (most recent call last)
in ()
—-> 1 fit = test.fit(X, Y)

~\Anaconda3\lib\site-packages\sklearn\feature_select'
339 Returns self.
340 """
–> 341 X, y = check_X_y(X, y, ['csr', 'csc'], multi_outpu
342
343 if not callable(self.score_func):

~\Anaconda3\lib\site-packages\sklearn\utils\validation
copy, force_all_finite, ensure_2d, allow_nd, multi_outp
y_numeric, warn_on_dtype, estimator)
571 X = check_array(X, accept_sparse, dtype, order, c
572 ensure_2d, allow_nd, ensure_min_samples,
–> 573 ensure_min_features, warn_on_dtype, estimato

574 if multi_output:

575 y = check_array(y, 'csr', force_all_finite=True, ensure_2d=False,

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in check_array(array, accept_sparse, dtype, order, copy, force_all_finite, ensure_2d, allow_nd, ensure_min_samples, ensure_min_features, warn_on_dtype, estimator)

431 force_all_finite)

432 else:

–> 433 array = np.array(array, dtype=dtype, order=order, copy=copy)

434

435 if ensure_2d:

ValueError: could not convert string to float: 'no'

can you guide me in this regard

Jason Brownlee November 17, 2017 at 9:19 am #

You may want to use a label encoder and a one hot encoder to convert string data to numbers.

Vinod P November 17, 2017 at 12:19 am #

import numpy as np
from pandas import read_csv
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
# load data
data = read_csv('C:\\Users\\abc\\Downloads\\xyz\\api.csv',names = ['org.apache.http.impl.client.DefaultHttpClient.execute','org.apache.http.impl.client.DefaultHttpClient.','java.net.URLConnection.getInputStream','java.net.URLConnection.connect','java.net.URL.openStream','java.net.URL.openConnection','java.net.URL.getContent','java.net.Socket.','java.net.ServerSocket.bind','java.net.ServerSocket.','java.net.HttpURLConnection.connect','java.net.DatagramSocket.','android.widget.VideoView.stopPlayback','android.widget.VideoView.start','and VideoView.setVideoPath','android.widget.VideoView.p 'android.text.format.DateFormat.getTimeFormat','andr telephony.TelephonyManager.listen','android.telephony ony.TelephonyManager.getSimSerialNumber','android. id.telephony.TelephonyManager.getLine1Number','and oid.speech.tts.TextToSpeech.','android.provider.Settin em.getInt','android.provider.Settings$System.getConfi ','android.provider.Settings$Secure.getInt','android.os os.PowerManager$WakeLock.release','android.os.Pov Manager.setWifiEnabled','android.net.wifi.WifiManage tate','android.net.wifi.WifiManager.getScanResults','ar d.media.RingtoneManager.getRingtone','android.medi

udioSource','android.media.MediaPlayer.stop','android.media.MediaPlayer.start','android.media.MediaPlay er.setDataSource','android.media.MediaPlayer.reset','android.media.MediaPlayer.release','android.media.M ediaPlayer.prepare','android.media.MediaPlayer.pause','android.media.MediaPlayer.create','android.media. AudioRecord.','android.location.LocationManager.requestLocationUpdates','android.location.LocationMana ger.removeUpdates','android.location.LocationManager.getProviders','android.location.LocationManager.ge tLastKnownLocation','android.location.LocationManager.getBestProvider','android.hardware.Camera.open' ,'android.bluetooth.BluetoothAdapter.getAddress','android.bluetooth.BluetoothAdapter.enable','android.blu etooth.BluetoothAdapter.disable','android.app.WallpaperManager.setBitmap','android.app.KeyguardManag e$KeyguardLock.reenableKeyguar','android.app.KeyguardManager$KeyguardLock.disableKeyguard','andr oid.app.ActivityManager.killBackgroundProcesses','android.app.ActivityManager.getRunningTasks','android .app.ActivityManager.getRecentTasks','android.accounts.AccountManager.getAccountsByType','android.ac counts.AccountManager.getAccounts','Class'])

```
dataframe = read_csv(url, names=names)
array = dataframe.values
X = array[:,0:70]
Y = array[:,70]
# feature extraction
model = LogisticRegression()
rfe = RFE(model, 3)
fit = rfe.fit(X, Y)
#print("Num Features: %d") % fit.n_features_
#print("Selected Features: %s") % fit.support_
#print("Feature Ranking: %s") % fit.ranking_
```

—————————————————————————————————————

I get following error
ValueError Traceback (most recent call last)
in ()
6 model = LogisticRegression()
7 rfe = RFE(model, 3)
—-> 8 fit = rfe.fit(X, Y)
9 print("Num Features: %d") % fit.n_features_
10 print("Selected Features: %s") % fit.support_

Jason Brownlee November 17, 2017 at 9:26 a

Perhaps try posting your code to stackov

Vinod P November 17, 2017 at 12:29 am #

Can you post a code on first select relevant fe
use relevant features to construct classification model'

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 17, 2017 at 9:26 am #   REPLY ↰

Thanks for the suggestion.

Hemalatha December 1, 2017 at 2:12 am #   REPLY ↰

will you post a code on selecting relevant features using feature selection method and then using relevant features constructing a classification model??

Jason Brownlee December 1, 2017 at 7:40 am #   REPLY ↰

Yes, see this post:

https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/

Arjun December 13, 2017 at 4:45 am #   REPLY ↰

Hi Jason,
Thanks for the content, it was really helpful.

Can you clarify if the above mentioned methods can also be used for regression models?

Jason Brownlee December 13, 2017 at 5:44 am #   REPLY ↰

Perhaps, I'm no sure off hand. Try and let me know how you go.

Danilo December 25, 2017 at 1:36 am #

Hi Jason,

I just had the same question as Arjun, I tried w
approaches were able to to do it.

Jason Brownlee December 25, 201

What was the problem exactly?

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

Zee Gola January 29, 2018 at 2:30 pm #

REPLY ↩

Hi Jason! Can you please further explain what the vector does in the separateByClass method?

Jason Brownlee January 30, 2018 at 9:46 am #

REPLY ↩

Sorry, I don't follow?

Anas January 29, 2018 at 8:57 pm #

REPLY ↩

Hi Jason,

Thank you for the post, it was very useful.

I have a regression problem with one output variable y (0<=y<=100) and 6 input features (I think that they are non-correlated).
The number of observations (samples) is 36980.
I used Random Forest algorithm to fit the prediction model.
The mean absolute error obtained is about 7.

Do you advise me to make features selection or not in this case?
In other words, from which number of features, it is advised to make features selection?

Congratulations.

Jason Brownlee January 30, 2018 at 9:50 am #

REPLY ↩

Try a suite of methods, build models from the selected features and see if the models outperform those models that use all features.

Joseph February 18, 2018 at 2:00 pm #

Hello Jason,
First thanks for sharing.
I have question with regards to four automatic feature
the same dataset. Pima dataset with exception of feat
magnitude.

Do you need to do any kind of scaling if the feature's n
other? For example if we assume one feature let's say

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

named "test" had values in range of 100s. Will this affect which automatic selector you choose or do you need to do any additional pre-processing?

Jason Brownlee February 19, 2018 at 9:03 am #   REPLY ↩

The scale of features can impact feature selection methods, it really depends on the method.

If you're in doubt, consider normalizing the data before hand.

Eric Williamson March 29, 2018 at 3:30 pm #   REPLY ↩

Feature scaling should be included in the examples.

The Pima Indians onset of diabetes dataset contains features with a large mismatches in scale. The rankings produced by the code in this article are influenced by this, and thus are not accurate.

Jason Brownlee March 30, 2018 at 6:29 am #   REPLY ↩

Thanks for the suggestion Eric.

Joseph February 18, 2018 at 3:34 pm #   REPLY ↩

Hello Jason,
One more question:
I noticed that when you use three feature selectors: Univariate Selection, Feature Importance and RFE you get different result for three important features.

1. When using Univariate with k=3 chisquare you get
plas, test, and age as three important features. (glucos

2. When using Feature Importance using ExtraTreesCl
The score suggests the three important features are pl
weight(bmi), and age)

3. When you use RFE
RFE chose the top 3 features as preg, mass, and pedi.
pedigree test.

According your article below
https://machinelearningmastery.com/an-introduction-t

Univariate is filter method and I believe the RFE and Fe
All three selector have listed three important features.

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

set of features and not the most reliable? However, the two other methods don't have same top three features? Are some methods more reliable than others? Or does this come down to domain knowledge?

Jason Brownlee February 19, 2018 at 9:04 am #   REPLY ↩

Different methods will take a different "view" of the data.

There is no "best" view. My advice is to try building models from different views of the data and see which results in better skill. Even consider creating an ensemble of models created from different views of the data together.

Ranbeer February 28, 2018 at 7:42 pm #   REPLY ↩

Hi Jason,

I'm your fan. Your articles are great. Two questions on the topic of feature selection

1. Shouldn't you convert your categorical features to "categorical" first?
2. Don't we have to normalize numeric features

Before doing PCA or feature selection? In my case it is taking the feature with the max value as important feature.
And, not all methods produce the same result.

Any thoughts?

Cheers,
Ranbeer

Jason Brownlee March 1, 2018 at 6:12 am #   REPLY ↩

Yes, Python requires all features to be numerical. Sometimes it can benefit the model if we rescale the input data.

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.

itisha March 5, 2018 at 7:41 am #

hi jason sir,
your articles are very helpful.
i have a confusion regarding gridserachcv()
i am working on sentiment analyis and i have created

Email Address

i am using linear SVC and want to do grid search for fir

START MY EMAIL COURSE

C, fir the model on train data and then test on test dat

different group of features one by one so do i need to do gridserach again and again for each feature group?

Jason Brownlee March 6, 2018 at 6:07 am #

REPLY ↰

Perhaps, it really depends how sensitive the model is to your data.

Also, so much grid searching may lead to some overfitting, be careful.

Yaseen March 10, 2018 at 3:16 am #

REPLY ↰

Thank you Jason for gentle explanation.
The last part "# Feature Importance with Extra Trees Classifier".
It looks the result is different if we consider the higher scores?

Jason Brownlee March 10, 2018 at 6:34 am #

REPLY ↰

Sorry, what do you mean exactly?

Aouedi Ons March 12, 2018 at 6:57 am #

REPLY ↰

Hi
Sir why you use just 8 example and your dataset contain many example ??

Jason Brownlee March 12, 2018 at 2:23 pm #

REPLY ↰

Sorry, I don't follow. Perhaps you can try

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.

Unni Mana April 5, 2018 at 1:30 am #

Hi Jason,

Your articles are awesome . After going through this ar

Out of these 4 suggested techniques, which one I have

Why the O/P is different based on different feature sele

Thanks

Email Address

START MY EMAIL COURSE

Jason Brownlee April 5, 2018 at 6:12 am #

Try them all and see which results in a model with the most skill.

Bhanupraksh Vattikuti April 10, 2018 at 6:09 pm #

Dear Jason,

Thank you the article.
When I am trying to use Feature Importance I am encountering the following error.
Can you please help me with this.
File "C:/Users/bhanu/PycharmProjects/untitled3/test_cluster1.py", line 14, in
model.fit(X, Y)
File "C:\Users\bhanu\PycharmProjects\untitled3\venv\lib\site-packages\sklearn\ensemble\forest.py", line
247, in fit
X = check_array(X, accept_sparse="csc", dtype=DTYPE)
File "C:\Users\bhanu\PycharmProjects\untitled3\venv\lib\site-packages\sklearn\utils\validation.py", line
433, in check_array
array = np.array(array, dtype=dtype, order=order, copy=copy)
ValueError: could not convert string to float: 'neptune'

Jason Brownlee April 11, 2018 at 6:32 am #

Perhaps you are running on a different dataset? Where did 'neptune' come from?

Dhanunjaya Mitta April 16, 2018 at 5:12 pm #

Can I get more information about Univariate F
correlation etc.,

Jason Brownlee April 17, 2018 at 5:54 am #

Thanks for the suggestion.

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

Elizabeth May 1, 2018 at 5:21 am #

Hi Jason,

Thank you for the post, it was very useful.

I have a problem that is one-class classification and I would like to select features from the dataset, however, I see that the methods that are implemented need to specify the target but I do not have the target since the class of the training dataset is the same for all samples.

Where can I found some methods for feature selection for one-class classification?

Thanks!

Jason Brownlee May 1, 2018 at 5:36 am #   REPLY ↩

If the class is all the same, surely you don't need to predict it?

Elizabeth May 1, 2018 at 6:10 am #   REPLY ↩

Well, my dataset is related to anomaly detection. So the training set contains only the objects of one class (normal conditions) and in the test set, the file combines samples that are under normal conditions and data from anomaly conditions.

What I understand is that in feature selection techniques, the label information is frequently used for guiding the search for a good feature subset, but in one-class classification problems, all training data belong to only one class.

For that reason, I was looking for feature selection implementations for one-class classification.

Ann May 2, 2018 at 10:57 pm #   REPLY ↩

Thank you for the post, it was very useful for beginner.
I have a problem that is I use Feature Importance with Extra Trees Classifier and how can
I display feature name(plas,age,mass,….etc) in this sa

for example:
Feature ranking:
1. plas (0.11070069)
2. age (0.2213717)
3. mass(0.08824115)
…….

Thanks for your help.

……..

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 May 3, 2018 at 6:34 am #

REPLY ↰

You might have to write some custom code I think.

Ronald Martis May 15, 2019 at 5:10 am #

REPLY ↰

Use the following:
print(list(zip(names, model.feature_importances_)))
You get:
[('preg', 0.11289758476179099), ('plas', 0.23098096297414206), ('pres', 0.09989914623444449),
('skin', 0.08008405837625963), ('test', 0.07442780491152233), ('mass', 0.14140399156908312),
('pedi', 0.11808706393665534), ('age', 0.142219387236102)]

Jason Brownlee May 15, 2019 at 8:19 am #

REPLY ↰

Nice!

Yanyun Zou May 8, 2018 at 1:13 pm #

REPLY ↰

Hi Jason,
I tried Feature Importance method, but all the values of variables are above 0.05, so does it mean that all
the variables have little relation with the predicted value?

Jason Brownlee May 8, 2018 at 2:56 pm #

REPLY ↰

Perhaps try other feature selection methods, build models from each set of features and
double down on those views of the features that re~~~~~~~~~~~~~~~~~~~~~~

Saeed Ullah June 7, 2018 at 5:01 am #

Hello Jason,
Thanks for you great post, I have a question in feature
ISOMAP or any other Dimensionality Reduction techni
features/dimensions is best for our classification algori

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 7, 2018 at 6:34 am #

Try multiple configurations, build and evaluate a model for each, use the one that results in the best model skill score.

Musthafa June 27, 2018 at 6:23 pm #

Hi,
what to do when i have multiple categorical features like zipcode,class etc
should i hot encode them

Jason Brownlee June 28, 2018 at 6:15 am #

Some like zip code you could use a word embedding.

Others like class can be one hot encoded.

Aymen July 21, 2018 at 3:54 am #

Hi,

Iwhen we use univariate filter techniques like Pearson correlation, mutul information and so on. Do we need to apply the filter technique on training set not on the whole dataset??

Jason Brownlee July 21, 2018 at 6:38 am #

Perhaps just work with the training data.

william July 26, 2018 at 7:41 pm #

jason – i'm working on several feature selectic
input variables for a continuous output variable i'm tryi
sure they apply to my specific regression problem i'm
a regression problem, are there any feature selection n
variable prediction?

i.e. i have normalized my dataset that has 100+ catego
a continuous output variable…any suggestions?

✕

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

thanks in advance!

Jason Brownlee July 27, 2018 at 5:52 am #

RFE will work for classification or regression. It's a good place to start.

Also, correlation of inputs with the output is another excellent starting point.

---

meenal jain August 2, 2018 at 8:29 pm #

I read your post, it was very productive.

Can i use linear correlation coefficient between categorical and continuous variable for feature selection.

or please suggest me some other method for this type of dataset (ISCX -2012) in which target class is categorical and all other attributes are continuous.

Jason Brownlee August 3, 2018 at 6:00 am #

No.

Perhaps look into feature importance scores.

Folmer August 3, 2018 at 11:45 pm #

Jason,

I was wondering whether the parameters of the machine learning tool that is used during the feature selection step are of any importance. Since most websites that I have seen so far just use the default parameter configuration during this phase.

I understand that adding a grid search has the followin
-It increase the calculation time substantially. (really wh
-Hard to determine which produces better results, real
machine learning tool.

But still, is it worth it to investigate it and use multiple
machine learning tool?

My situation:
-A (non-linear) dataset with ~20 features.
-Planning to use XGBooster for the feature selection p
was sufficient).

-For the construction of the model I was planning to use MLP NN, using a gridsearch to optimize the parameters.

Thanks in advance!

Jason Brownlee August 4, 2018 at 6:12 am #

Yes you can tune them.

Generally, I recommend generating many different "views" on the inputs, fit a model to each and compare the performance of the resulting models. Even combine them.

Most likely, there is no one best set of features for your problem. There are many with varying skill/capability. Find a set or ensemble of sets that works best for your needs.

Sa August 13, 2018 at 9:14 pm #

Hi Jason

I need to do feature engineering on rows selection by specifying the best window size and frame size , do you have any example available online?

thanks
Sa

Jason Brownlee August 14, 2018 at 6:19 am #

For time series, yes right here:

https://machinelearningmastery.com/sensitivity-analysis-history-size-forecast-skill-arima-python/

Yvonne August 24, 2018 at 12:00 am #

Hi Jason

I am a beginner in python and scikit learn. I am current
and healthy controls based on functional connectivity
out of approximately 20'000 features. I get 32 selected
next is to run a permutation statistic to check if my res

My question: Do I have to run the permutation statistic
the 20'000 for this purpose.

Below you can see my code. to simplify my question, i
identical. I would appreciate your help very much, as I

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

Best Yolanda

homeDir = 'F:\Analysen\Prediction_TreatmentOutcome\PyCharmProject_TreatmentOutcome' # location of the connectivity matrices

```
# ################################################################################
# import packages needed for classification
import numpy as np
import os

import matplotlib.pyplot as plt
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_validate
from sklearn.model_selection import StratifiedKFold
from sklearn.feature_selection import RFECV
from sklearn import svm

from sklearn.pipeline import make_pipeline, Pipeline
from sklearn import preprocessing

from sklearn.model_selection import permutation_test_score

class PipelineRFE(Pipeline):

def fit(self, X, y=None, **fit_params):
super(PipelineRFE, self).fit(X, y, **fit_params)
self.coef_ = self.steps[-1][-1].coef_
return self

clf = PipelineRFE(
[
('std_scaler', preprocessing.StandardScaler()), #z-transformation
("svm", svm.SVC(kernel = 'linear', C = 1)) #estimator
]
)

# #######################################
# Load and prepare data set

nQuest = 5 # number of questionnaires

samples = np.loadtxt('FBDaten_T1.txt')

# Import targets (created in R based on group variable
targets =
np.genfromtxt(r'F:\Analysen\Prediction_TreatmentOutc
L_fbDaten.txt', dtype= str)

# #######################################
# run classification
```

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

```
skf = StratifiedKFold(n_splits = 5) # The folds are made by preserving the percentage of samples for each
class.

# rfecv
rfecv = RFECV(estimator = clf, step = 1, cv = skf, scoring = 'accuracy')
rfecv.fit(samples, targets)

# The number of selected features with cross-validation.
print("Optimal number of features : %d" % rfecv.n_features_)

# Plot number of features VS. cross-validation scores
plt.figure()
plt.xlabel("Subset of features")
plt.ylabel("Cross validation score (nb of correct classifications)")
plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
plt.show()

#The mask of selected features
rfecv.support_
print("Mask of selected features : %s" % rfecv.support_)

#Find index of true values in a boolean vector
index_features = np.where(rfecv.support_)[0]
print(index_features)

#Find value of indices
reduced_features = samples[:, index_features]
print(reduced_features)

## permutation testing on reduced features

score, permutation_scores, pvalue = permutation_test_score(
clf, reduced_features, targets, scoring="accuracy", cv=skf, n_permutations=100, n_jobs=1)

print("Classification score %s (pvalue : %s)" % (score, pvalue))
```

Jason Brownlee August 24, 2018 at 6:11 am   REPLY ↰

Sorry, I do not have the capacity to review

**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.

Rafi August 29, 2018 at 7:00 am #

Thank you a lot for this useful tutorial. It would
cons/pros of each method.
Thanks in advance.

Email Address

START MY EMAIL COURSE

Jason Brownlee August 29, 2018 at 8:18 am #   REPLY ↩

Thanks for the suggestion.

Mohammed August 30, 2018 at 8:42 pm #   REPLY ↩

I want to ask about feature extraction procedure, what's the criteria to stop training and extract features. Are it depend on the test accuracy of model?. In other meaning what is the difference between extract feature after train one epoch or train 100 epoch? what is best features?, may be my question foolish but i need answer for it.

Jason Brownlee August 31, 2018 at 8:10 am #   REPLY ↩

What do you mean by extract features? Here, we are talking about feature selection?

Mohammed August 30, 2018 at 8:51 pm #   REPLY ↩

I ask about feature extraction procedure, for example if i train CNN, after which number of epochs should stop training and extract features?. In other meaning are feature extraction depend on the test accuracy of training model?. If i build model (any deep learning method) to only extract features can i run it for one epoch and extract features?

Jason Brownlee August 31, 2018 at 8:10 am #   REPLY ↩

I see.

You want to use features from a model that is skillful. Perhaps at the same task, perhaps at a reconstruction task (e.g. an autoencoder).

Mohammed August 31, 2018 at 9:32 pm #

I do not understand answer

Jason Brownlee September 1, 2018

Sorry, which part?

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

munaza August 30, 2018 at 9:55 pm #

Hello Sir,

Thank you soo much for this great work.

Will you please explain how the highest scores are for : plas, test, mass and age in Univariate Selection. I am not getting your point.

Jason Brownlee August 31, 2018 at 8:11 am #

What problem are you having exactly?

munaza August 31, 2018 at 2:23 pm #

Thank you sir for the reply…

Actually I was not able to understand the output of chi^2 for feature selection. The problem has been solved now.

Thanks a lot.

Jason Brownlee September 1, 2018 at 6:15 am #

I'm happy to hear that you solved your problem.

Gabriel Joshua Migue September 4, 2018 at 10:

Which is the best technique for feature select

change when i try multiple times?

Jason Brownlee September 5, 2018 at 6:40 a

This is a common question that I answer

https://machinelearningmastery.com/faq/single-fa

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

Paul September 15, 2018 at 8:33 pm #

Hi,

in your example for feature importance you use as Ensemble classifier the ExtraTreesClassifier.
In sci-kit learn the default value for bootstrap sample is false.

Doesn't this contradict to find the feature importance? e.g it could build the tree on only one feature and so the importance would be high but does not represent the whole dataset.

Thanks
Paul

Jason Brownlee September 16, 2018 at 5:59 am #

Trees will sample features and in aggregate the most used features will be "important".

It only means the features are important to building trees, you can interpret it how ever you like.

Abdur Rehmna September 21, 2018 at 4:38 am #

Hi Jason,

I have a dataset which contains both categorical and numerical features. Should I do feature selection before one-hot encoding of categorical features or after that ?

Jason Brownlee September 21, 2018 at 6:34 am #

Sure. It's a cheap operation (easy) and has big effects on performance.

Dean September 23, 2018 at 7:41 am #

Hi Jason

I haven't read all the comments, so I don't know if this
this:

https://hub.packtpub.com/4-ways-implement-feature-

It's identical (barring edits, perhaps) to your post here,
I should let you know.

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 24, 2018 at 6:06 am # REPLY ↩

That is very disappointing!

Thanks for letting me know Dean.

zhyar September 30, 2018 at 12:31 am # REPLY ↩

Hi Jason

I have a dataset with two classes. In the feature selection, I want to specify important features for each class. For example, if I chose 15 important features, determine which attribute is more important for which class.please help me

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

Yes, this is what feature selection will do for you.

Thiago Batista Soares October 1, 2018 at 6:05 am # REPLY ↩

Hi Jason

First, congratulations on your posts and your books.
I am reaing your book machine learning mastery with python and chapter 8 is about this topic and I have a doubt, should I use thoses technical with crude data or should I normalize data first? I did test both case but results are different, exemple (first case column A and B are important but second case column C and D are important)

Very thanks.

Jason Brownlee October 1, 2018 at 6:33 am

Thanks.

Build models from each and go with the approach
hold out dataset.

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

isuru dilantha October 24, 2018 at 1:54 pm #

HI Jason,

I'm on a project to predict next movement of animals using their past data like location, date and time. what are the possible models that i can use to predict their next location ?

Jason Brownlee October 24, 2018 at 2:50 pm #

I recommend following this process for new problems:

https://machinelearningmastery.com/start-here/#process

Hannes November 7, 2018 at 12:02 am #

hi,

Many thanks for your hard work on explaining ML to the masses.

I'm trying to optimize my Kaggle-kernel at the moment and I would like to use feature selection. Because my source data contains NaN, I'm forced to use an imputer before the feature selection.

Unfortunately, that results in actually worse MAE then without feature selection.

Do you have a tip how to implement a feature selection with NaN in the source data?

Jason Brownlee November 7, 2018 at 6:07 am #

Perhaps you can remove the rows with NaNs from the data used to train the feature selector?

Supr November 28, 2018 at 7:24 am #

Hi Jason,
Somehow ur blog almost always has exactly what I need. Least I could do is say thanks and wish u all the best!

Jason Brownlee November 28, 2018 at 7:48 a

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

PC December 13, 2018 at 1:06 pm #

Hi Jason,

Your work is amazing. Got interested in Machine learning after visiting your site. Thank You, Keep up your good work.

I tried using RFE in another dataset in which I converted all categorical values to numerical values using Label Encoder but still I get the following error:

————————————————————————

ValueError Traceback (most recent call last)

in ()

14 model = LogisticRegression()

15 rfe = RFE(model, 5)

—> 16 fit = rfe.fit(X, Y)

17 print("Num Features: %d" % fit.n_features_)

18 print("Selected Features: %s" % fit.support_)

~\Anaconda3\lib\site-packages\sklearn\feature_selection\rfe.py in fit(self, X, y)

132 The target values.

133 """"

–> 134 return self._fit(X, y)

135

136 def _fit(self, X, y, step_score=None):

~\Anaconda3\lib\site-packages\sklearn\feature_selection\rfe.py in _fit(self, X, y, step_score)

140 # self.scores_ will not be calculated when calling _fit through fit

141

–> 142 X, y = check_X_y(X, y, "csc")

143 # Initialization

144 n_features = X.shape[1]

~\Anaconda3\lib\site-packages\sklearn\utils\validation.py in check_X_y(X, y, accept_sparse, dtype, order, copy, force_all_finite, ensure_2d, allow_nd, multi_output, ensure_min_samples, ensure_min_features, y_numeric, warn_on_dtype, estimator)

571 X = check_array(X, accept_sparse, dtype, order, copy, force_all_finite,

572 ensure_2d, allow_nd, ensure_min_samples,

–> 573 ensure_min_features, warn_on_dtype, estimator)

574 if multi_output:

575 y = check_array(y, 'csr', force_all_finite=True, ensu

~\Anaconda3\lib\site-packages\sklearn\utils\validatio
order, copy, force_all_finite, ensure_2d, allow_nd, ensu
warn_on_dtype, estimator)

431 force_all_finite)

432 else:

–> 433 array = np.array(array, dtype=dtype, order=ord

434

435 if ensure_2d:

ValueError: could not convert string to float: 'StudentA

I am in dire need of a solution for this. Kindly help me .

REPLY ↩

Jason Brownlee December 13, 2018 at 1:49 pm #

It suggests your data file may still have string values.

Perhaps double check your loaded data?

REPLY ↩

PC December 13, 2018 at 4:27 pm #

I had checked the data type of that particular column and it is of type int64 as given below:

In:

mod_StudentData['StudentAbsenceDays'].dtype

Out[]:
dtype('int64')

REPLY ↩

Jason Brownlee December 14, 2018 at 5:29 am #

Nice work!

REPLY ↩

Oswaldo Castro December 15, 2018 at 4:54 am #

Hi Jason…
Great article as usual.
I'm novice in ML and the article leaves me with a doubt.
The SelectKBest, RFE and ExtraTreesClassifier are performing Feature Extraction and PCA is performing
Feature Extraction.
Am I right with this?
Thanks Jason

Jason Brownlee December 15, 2018 at 6:16 a

Yes.

afer December 24, 2018 at 10:49 am #

Hi Jason,
First of all nice tutorial!
My question is that I have a samples of around 30,000 with around 150 features each for a binary classification problem. The plan is to do RFE on GRID SEARCH (Select features and tune parameters at the same pipeline) using a 3-fold cross validation (Each fold, the data is split twice, one for RFE and another for GRID SEARCH), this is done on the entire data set.

Now, after determining the best features and parameters, using the SAME data set, I split it into training / validation / test set and train a model using the selected features and parameters to obtain its accuracy (of the best model possible, and on the test set, of course).

Is this the correct thing to do? My reason for this methodology is that, the feature/parameter selection is a whole different process from the actual model fitting (using the selected features and parameters), meaning the actual model fitting will not actually know what the feature/parameter selection learned on the entire dataset, hence it is only okay to re-use the entire data set.

If this is not the case, what would you recommend? perhaps, separate the entire data set into a feature/parameter selection set and actual model fitting set (50:50), wherein after the best features and parameters have been determined on the first 50%, use these features on the remaining 50% of the data to train a model (this 50 is further split into train/validation/test).

Jason Brownlee December 25, 2018 at 7:17 am #

I recommend performing feature selection on each fold of CV or with a separate dataset up front.

Aaron Carl Fernandez January 3, 2019 at 2:11 pm #

Thank you for the answer Dr. Jason! Also, the grid selection + RFE process is going to spit out the accuracy / F1-score of the best model this be considered as the FINAL score of the y build another model (the final model with your score of the model's performance?

Jason Brownlee January 4, 2019 a

I recommend building a final mod harness may be a suitable estimate of mo on your project regarding what is satisfact

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

Aaron Carl Fernandez January 8, 2019 at 1:52 am #

Thanks Dr. Jason. One last question :), can I use the chi squared statistical test (in the univariate selection portion) for reflecting the p-value or the statistical significance of each feature? Let say, I am going to show the trimmed mean of each feature in my data, does the chi squared p-value confirm the statistical significance of the trimmed means?

Jason Brownlee January 8, 2019 at 6:52 am #

No, it comments on the relationship between categorical variables.

REPLY

Aaron Carl Fernandez January 10, 2019 at 12:40 am #

Thanks Dr. Jason!.. One last question promise 🙂 I assume its okay to prune my features and parameters on a 3-fold nested cross-validated RFE + GS while building my final model on a 10-fold regular cross validation. I used different data sets on each process (I split the original dataset 50:50, used the first half for RFE + GS and the second half to build my final model). The reason is that the nested cross-validated RFE + GS is too computationally expensive and that I'd like to train my final model on a finer granularity hence, the regular 10-fold CV.

Thank you soo much!

REPLY

Jason Brownlee January 10, 2019 at 7:52 am #

I cannot comment if your test methodology is okay, you must evaluate it in terms of stability/variance and use it if you feel the results will be reliable.

Aaron Carl Fernandez January 10, 20

Thank you so much Dr. Jason

Houssam February 4, 2019 at 11:03 pm #

Hi Dr. Jason;
I want to ask you a question: I want to apply the PSO
diabetes dataset, I am disturbed, what can I do

Jason Brownlee February 5, 2019 at 8:23 am #   REPLY ↩

Sorry, I don't have examples of using global optimization algorithms for feature selection – I'm not convinced that the techniques are relatively effective.

maedeh February 12, 2019 at 5:42 pm #   REPLY ↩

Hi,
I am a beginner and my question may be wrong. can we use these feature selection methods in an autoencoder that our inputs and outputs of our network are an image for example mnist? Thanks

Jason Brownlee February 13, 2019 at 7:54 am #   REPLY ↩

Not really, you would be performing feature selection on pixel values.

The autoencoder is doing a form of this for you.

Rahul Ramaswamy February 20, 2019 at 7:36 pm #   REPLY ↩

Hi Jason,
All the techniques mentioned by you works perfectly if there is a target variable (Y or 8th column in your case). The dataset i am working on uses unsupervised learning algorithm and hence does not have any target/dependent variable. Does the feature selection work in such cases? If yes, how should i go about it.

Jason Brownlee February 21, 2019 at 7:54 am #   REPLY ↩

It can, but you may have to use a method information or correlation, etc.

I don't have a worked example, sorry.

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

Selene March 4, 2019 at 9:19 am #

Hello Jason,
Your blog and the way how you explain things are fant
real applications, the fit method of some feature select

set or to the whole data set (training + testing)?

Thanks a lot!

Jason Brownlee March 4, 2019 at 2:17 pm #   REPLY ↩

It is fit on just the training dataset when evaluating a model. It is fit on all data when developing a final model:

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

Selene March 7, 2019 at 12:05 am #   REPLY ↩

Thank you very much!

Keerti Bafna March 7, 2019 at 5:36 am #   REPLY ↩

Hello Jason,

First of all thank you for such an informative article.

I need to perform a feature selection using the Filter, Wrapper and Embedded methods. The plan is to then take an average of scores from each selection procedure and select the top 10 features.

My plan is to split the data initially into train and hold out sets. I plan to then use cross-validation for each of the above 3 methods and use only the train data for this (internally in each of the fold). Once i get my top 10 features , i will then only use them in the hold out set and predict my model performance.

Do you feel this method would give me a stable model ? If not, what can i improve / change ?

Thanks in advance.

Jason Brownlee March 7, 2019 at 7:00 am #

Perhaps try it and see.

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.

Mohamed Saad March 10, 2019 at 5:07 pm #

Hi Jason,
thank you about your efforts,
I want to ask about Feature Extraction with RFE

Email Address

START MY EMAIL COURSE

I use your mention code
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
and the results are :
[1 2 3 5 6 1 1 4]

when I change the order of columns names as I mention

names = ['pedi','preg', 'plas', 'pres', 'test', 'age', 'class','mass','skin']
I get same output
[1 2 3 5 6 1 1 4]

can you explain how it work?
thank you

Jason Brownlee March 11, 2019 at 6:48 am # REPLY ↩

Changing the order of the labels does not change the order of the columns in the dataset. This is why you are getting the same output indexes.

Mohamed Saad March 11, 2019 at 11:35 am # REPLY ↩

I try to change the order of columns to check the validity of the RFE rank.
what is your advice if I want to check the validity of rank?

Also, I want to ask when I try to choose the features that influence on my models, I should add all features in my dataset (numerical and categorical) or only categorical features?

Thank you

Jason Brownlee March 11, 2019 at 2:14 pm # REPLY ↩

I'm not sure this is required.

Nevertheless, you would have to change the colum
CSV from which it was loaded.

It depends on the capabilities of the feature select
selection.

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

Waqar April 3, 2019 at 8:03 am #

I am unable to get output, because of this wa

"C:\Users\Waqar\Anaconda3\lib\site-packages\sklearn\model_selection_split.py:626: Warning: The least populated class in y has only 1 members, which is too few. The minimum number of members in any class cannot be less than n_splits=5."

Please help if someone knows.

Jason Brownlee April 3, 2019 at 4:11 pm #   REPLY ↩

Perhaps you can use fewer splits or use more data?

Viva April 13, 2019 at 6:39 pm #   REPLY ↩

Hi Jason,

I am reading from your book on ML Mastery with Python and I was going to the same topic mentioned above, I see you have chose chi square to do feature selection in univariate method, how do I decide to choose between different tests (chi square, t-test , ANOVA).

Jason Brownlee April 14, 2019 at 5:46 am #   REPLY ↩

Great question. I answer it here:

https://machinelearningmastery.com/faq/single-faq/what-feature-selection-method-should-i-use

---

Viva April 15, 2019 at 7:15 pm #   REPLY ↩

Thank you, a big post to read for next learning steps 🙂

Jason Brownlee April 16, 2019 at 6:47 am #   REPLY ↩

Thanks.

Upender Singh June 7, 2019 at 6:23 pm #

hi jason

please let me how to choose best k value in case of us

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 8, 2019 at 6:49 am #

Evaluate models for different values of k and choose the value for k that gives the most skillful model.

shruthi June 16, 2019 at 4:09 pm #

hi jason,

i want to use Univariate selection method.
I am building a linear regression model which has around 46 categorical variables.
if i want to know the best categorical features to be used in building my model, I need to send onehot encoded values to fit function right and the score_func sjould also be chi2 ?

test = SelectKBest(score_func=chi2, k=4)
fit = test.fit(X, Y)

In the above code X should be the one hot encoded values of all the categorical variables right ?

Thanks in advance

Jason Brownlee June 17, 2019 at 8:16 am #

Perhaps this will help:

https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

Also, linear regression sounds like a bad fit, try a decision tree, and some other algorithms as well.

Rushali Jaiswal June 18, 2019 at 3:31 am #

Hi,
I had a question.
Do you apply feature selection before creating the dun

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.

Email Address

START MY EMAIL COURSE

Jason Brownlee June 18, 2019 at 6:42 am #

Feature selection is performed before.

Leave a Reply

Name (required)

Email (will not be published) (required)

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.
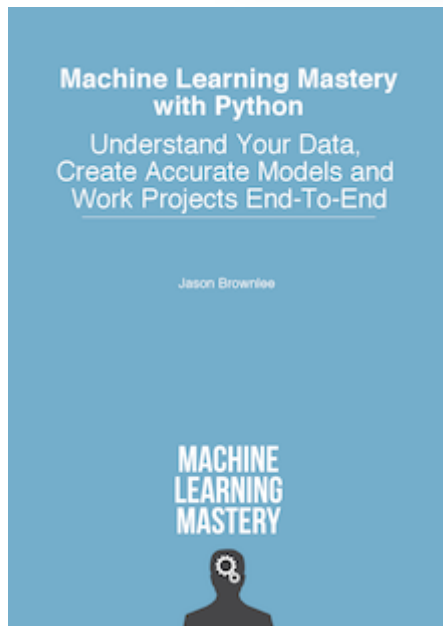
Read More

Machine Learni

Develop predictive mo

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

POPULAR

How to Develop LSTM Models for Multi-Step Time Series Forecasting of Household Power Consumption
OCTOBER 10, 2018

How to Develop LSTM Models for Time Series Forecasting
NOVEMBER 14, 2018

11 Classical Time Series Forecasting Methods in Python (Cheat Sheet)
AUGUST 6, 2018

A Gentle Introduction to LSTM Autoencoders
NOVEMBER 5, 2018

How to Develop RNN Models for Human Activity Recognition Time Series Classification
SEPTEMBER 24, 2018

How to Develop Convolutional Neural Network Mod
NOVEMBER 12, 2018

When to Use MLP, CNN, and RNN Neural Networks
JULY 23, 2018

Top beginner tutorials:

- How to Install Python for Machine Learning

- Your First Machine Learning Project in Python
- Your First Neural Network in Python
- Your First Classifier in Weka
- Your First Time Series Forecasting Project

Top crash courses:

- Python Machine Learning
- Deep Learning for Computer Vision
- Deep Learning for Time Series
- Linear Algebra
- Statistical Methods

Top project tutorials:

- Face Detection Project
- Object Detection Project
- Photo Captioning Project
- Machine Translation Project
- Sentiment Analysis Project
- Power Forecasting Project

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