An Introduction to Feature Selection

by **Jason Brownlee** on October 6, 2014 in **Data Preparation**

Last Updated on June 29, 2021

Which features should you use to create a predictive model?

This is a difficult question that may require deep knowledge of the problem domain.

It is possible to automatically select those features in your data that are most useful or most relevant for the problem you are working on. This is a process called feature selection.

In this post you will discover feature selection, the types of methods that you can use and a handy checklist that you can follow the next time that you need to select features for a machine learning model.

Kick-start your project with my new book Data Preparation for Machine Learning, including *step-by-step tutorials* and the *Python source code* files for all examples.

Let's get started.

An Introduction to
Photo by John Tann,

What is Feature Selection

Feature selection is also called variable selection or a

It is the automatic selection of attributes in your data (
relevant to the predictive modeling problem you are w

> *feature selection… is the process of selecting a subset of relevant features for use in model construction*

— Feature Selection, Wikipedia entry.

Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, where as feature selection methods include and exclude attributes present in the data without changing them.

Examples of dimensionality reduction methods include Principal Component Analysis, Singular Value Decomposition and Sammon's Mapping.

> *Feature selection is itself useful, but it mostly acts as a filter, muting out features that aren't useful in addition to your existing features.*

— Robert Neuhaus, in answer to "How valuable do yo

The Problem The Feature Selec

Feature selection methods aid you in your mission to c u by choosing features that will give you as good or bett

Feature selection methods can be used to identify and attributes from data that do not contribute to the accur the accuracy of the model.

Fewer attributes is desirable because it reduces the c simpler to understand and explain.

> *The objective of variable selection is three-fold predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.*

— Guyon and Elisseeff in "An Introduction to Variable and Feature Selection" (PDF)

Feature Selection Algorithms

There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods.

Filter Methods

Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable.

Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores.

Wrapper Methods

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy.

The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features.

An example if a wrapper method is the recursive feature elimination algorithm.

Embedded Methods

Embedded methods learn which features best contribu s being created. The most common type of embedded f ods.

Regularization methods are also called penalization m ne optimization of a predictive algorithm (such as a regre complexity (fewer coefficients).

Examples of regularization algorithms are the LASSO

Feature Selection Tutorials and

We have seen a number of examples of features sele

- **Weka**: For a tutorial showing how to perform featu o Improve Accuracy and Decrease Training Time".
- **Scikit-Learn**: For a recipe of Recursive Feature Elimination in Python using scikit-learn, see "Feature Selection in Python with Scikit-Learn".
- **R**: For a recipe of Recursive Feature Elimination using the Caret R package, see "Feature Selection with the Caret R Package"

A Trap When Selecting Features

Feature selection is another key part of the applied machine learning process, like model selection. You cannot fire and forget.

It is important to consider feature selection a part of the model selection process. If you do not, you may inadvertently introduce bias into your models which can result in overfitting.

> *… should do feature selection on a different dataset than you train [your predictive model] on … the effect of not doing this is you will overfit your training data.*

— Ben Allison in answer to "Is using the same data for feature selection and cross-validation biased or not?"

For example, you must include feature selection within the inner-loop when you are using accuracy estimation methods such as cross-validation. This means that feature selection is performed on the prepared fold right before the model is trained. A mistake would be to perform feature selection first to prepare your data, then perform model selection and training on the selected features.

> *If we adopt the proper procedure, and perform feature selection in each fold, there is no longer any information about the held out cases in the choice of features used in that fold.*

— Dikran Marsupial in answer to "Feature selection fo
machine learning"

The reason is that the decisions made to select the fe
turn are passed onto the model. This may cause a mo
over other models being tested to get seemingly bette

> *If you perform feature selection on all of the da
each fold of the cross-validation procedure wa
what biases the performance analysis.*

— Dikran Marsupial in answer to "Feature selection ar

Feature Selection Checklist

Isabelle Guyon and Andre Elisseeff the authors of "An Introduction to Variable and Feature Selection" (PDF) provide an excellent checklist that you can use the next time you need to select data features for you predictive modeling problem.

I have reproduced the salient parts of the checklist here:

1. **Do you have domain knowledge?** If yes, construct a better set of ad hoc"" features
2. **Are your features commensurate?** If no, consider normalizing them.
3. **Do you suspect interdependence of features?** If yes, expand your feature set by constructing conjunctive features or products of features, as much as your computer resources allow you.
4. **Do you need to prune the input variables (e.g. for cost, speed or data understanding reasons)?** If no, construct disjunctive features or weighted sums of feature
5. **Do you need to assess features individually (e.g. to understand their influence on the system or because their number is so large that you need to do a first filtering)?** If yes, use a variable ranking method; else, do it anyway to get baseline results.
6. **Do you need a predictor?** If no, stop

7. **Do you suspect your data is "dirty" (has a few meaningless input patterns and/or noisy outputs or wrong class labels)?** If yes, detect the outlier examples using the top ranking variables obtained in step 5 as representation; check and/or discard them.
8. **Do you know what to try first?** If no, use a linear predictor. Use a forward selection method with the "probe" method as a stopping criterion or use the 0-norm embedded method for comparison, following the ranking of step 5, construct a sequence of predictors of same nature using increasing subsets of features. Can you match or improve performance with a smaller subset? If yes, try a non-linear predictor with that subset.
9. **Do you have new ideas, time, computational resources, and enough examples?** If yes, compare several feature selection methods, including your new idea, correlation coefficients, backward selection and embedded methods. Use linear and non-linear predictors. Select the best approach with model selection
10. **Do you want a stable solution (to improve per** subsample your data and redo your analysis for s

Further Reading

Do you need help with feature selection on a specific p u started fast:

- How to perform feature selection in Weka (withou
- How to perform feature selection in Python with s
- How to perform feature selection in R with caret

To go deeper into the topic, you could pick up a dedica g:

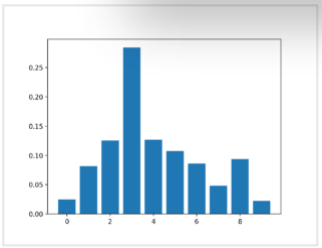
- Feature Selection for Knowledge Discovery and Data Mining
- Computational Methods of Feature Selection
- Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches
- Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop
- Feature Extraction, Construction and Selection: A Data Mining Perspective

You might like to take a deeper look at feature engineering in the post:

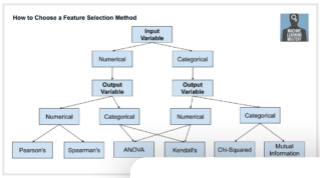
- Discover Feature Engineering, How to Engineer Features and How to Get Good at It

More On This Topic

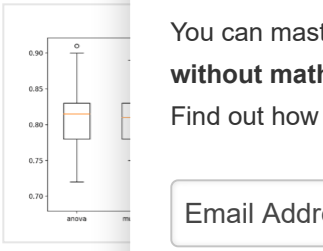
Start Machine Learning



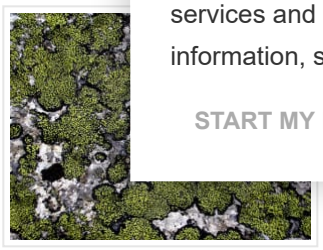
How to Calculate Feature Importance With Python



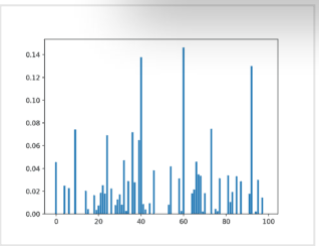
How to Choose a Feature S



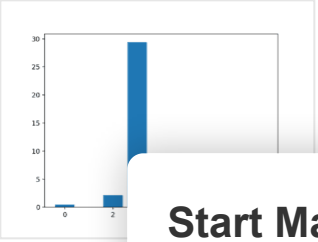
How to Develop a Feature S



Discover Feature Engineering, How to Engineer…



How to Perform Feature Selection for Regression Data



How to Perform Feature S

About Jason Brownlee

Jason Brownlee, PhD is a machine learnin th modern machine learning methods via han

View all posts by Jason Brownlee →

‹ Inteview: Discover the Methodology and Mindset of a Kaggle M ure ›

224 Responses to *An Introduction to Feature Selection*

Zvi Boger October 2, 2015 at 12:05 am #

REPLY

People can use my automatic feature dimension reduction algorithm published in:

Z. Boger and H. Guterman, Knowledge extraction from artificial neural networks models. Proceedings of the IEEE International Conference on Systems Man and Cybernetics, SMC'97, Orlando, Florida, Oct. 1997, pp. 3030-3035.

or contact me at optimal@peeron.com to get a copy of the paper..

The algorithm analyzes the "activities" of the trained model's hidden neurons outputs. If a feature dose not contribute to these activities, it either "flat" in the data, or the connection weights assigned to it are too small.

In both cases it can be safely discarded and the ANN retrained with the reduced dimensions.

Jason Brownlee December 29, 2015 at 4:13 pm #

REPLY

Thanks for sharing Zvi.

Joseph December 29, 2015 at 2:38 pm #

REPLY

Nice Post Jason, This is an eye opener for me and I have been looking for this for quite a while. But my challenge is quite different I think, my dataset is still in raw form and comprises different relational tables. How to select best features and how to form a new matrix for my predictive modelling are the major challenges I am facing.

Thanks

Jason Brownlee December 29, 2015 at 4:12

Thanks Joseph.

I wonder if you might get more out of the post on fe

doug February 9, 2016 at 4:22 pm #

very nice synthesis of some of the 'primary sc

Jason Brownlee July 20, 2016 at 5:26 am #

REPLY

Thanks doug.

bura February 9, 2016 at 4:58 pm #

REPLY

hello
Can we use selection teqnique for the best features in the dataset that is value numbering?

Jason Brownlee July 20, 2016 at 5:27 am #

REPLY

Hi bura, if you mean integer values, then yes you can.

swati March 6, 2016 at 10:07 pm #

REPLY

how can chi statistics feature selection algorithm work in data reduction.

Jason Brownlee July 20, 2016 at 5:31 am #

REPLY

The calculated chi-squared statistic can be used within a filter selection method.

Poornima July 21, 2017 at 2:40 pm #

Which is the best tool for chi square f

Jason Brownlee July 22, 2017 at 8

It is supported on most platforms.

Poornima July 24, 2017 at 3:36 p

Actually i want to apply Chi sq
to find the redundancy between the tw
only compute the level of independence between the attribute and the class attribute. May question is…what is the exact formula to use Chi square to find the level of independence between two attributes….? PS: I cannot use an existing tool…thanks

Jason Brownlee July 25, 2017 at 9:32 am #

Sorry, I don't have the formula at hand. I'd recommend checking a good stats text or perhaps Wikipedia.
https://en.wikipedia.org/wiki/Chi-squared_test

Ralf May 2, 2016 at 5:56 pm #

REPLY

which category does Random Forest's feature importance criterion belong as a feature selection technique?

Jason Brownlee July 20, 2016 at 5:29 am #

REPLY

Great question Ralf.

Relative feature importance scores from RandomForest and Gradient Boosting can be used as within a filter method.

If the scores are normalized between 0-1, a cut-off can be specified for the importance scores when filtering.

yerart July 25, 2019 at 12:25 am #

REPLY

Well, Jason and Ralf, I would first thin
perform the 'feature selection as part of the alg
can use their variable importance by-product a
approach of feature selection. But as a filter …
any other (statistical?) method other than using
(searched or whatnot) chosen subset of featur

Jason Brownlee July 25, 2019 at 7

Yes, the distinction is less clear.

I would treat feature importance scores fro

swati June 23, 2016 at 10:58 pm #

REPLY

CHI feature selection ALGORITHM IS is NP- HARD OR NP-COMPLETE

Jason Brownlee July 20, 2016 at 5:29 am #

REPLY

I'm not sure swati, but does it matter?

Vicky September 11, 2019 at 5:42 pm #

REPLY

Hi Jason i hope you are doing well, thanks a lot for the post. I am new to learn about feature selection. What you suggest me for the beginners

Jason Brownlee September 12, 2019 at 5:15 am #

REPLY

Good question, I recommend RFE.

Waqar Khan September 12, 2019 at 11:59 am #

Thanks Jason

Mohammed AbdelAal June 26, 2016 at 9:53 pm

Hi all,
Thanks Jason Brownlee for this wonderful article.

I have a small question. While performing feature selec
the feature selection method selects NO features?. Do

Jason Brownlee June 27, 2016 at 5:42 am #

Good question. If this happens, you will n
like a good one to me.

Dado July 19, 2016 at 10:20 pm #

REPLY

Hello Jason!

Great site and great article. I'm confused about how the feature selection methods are categorized though:

Do filter methods always perform ranking? Is it not possible for them to use some sort of subset search strategy such as 'sequential forward selection' or 'best first'?'

Is it not possible for wrapper or embedded methods to perform ranking? For example when I select 'Linear SVM' or "LASSO" as the estimator in sklearns 'SelectFromModel'-function, it seems to me that it examines each feature individually. The documentation doesn't mention anything about a search strategy.

Jason Brownlee July 20, 2016 at 5:34 am #

REPLY

Good question Dado.

Start Machine Learning

Feature subsets can be created and evaluated using a technique in the wrapper method, this would not be a filter method.

You can use an embedded within a wrapper method, but I expect the results would be less insightful.

Youssef August 9, 2016 at 7:09 pm #

REPLY

Hi, thx all or your sharing
I had a quation about the limitation of these methods in terms of number of features. In my scope we work on small sample size (n=20 to 40) with a lot of features (up to 50)
some people suggested to do all combinations to get high performence in terms of prediction.
what do you think?

Jason Brownlee August 15, 2016 at 11:14 am

I think try lots of techniques and see what

Jarking August 9, 2016 at 9:28 pm #

hi,I'm now learning feature selection with hier
begin with it?could you give me some ideas?

Jason Brownlee August 15, 2016 at 11:15 am #

REPLY

Consider starting with some off the shelf techniques first.

L K September 3, 2016 at 3:06 am #

REPLY

hi,
i want to use feature extractor for detecting metals in food products through features such as amplitude and phase. Which algorithm or filter will be best suited?

Jason Brownlee September 3, 2016 at 6:59 am #

REPLY

Here is a tutorial for feature selection in Python that may give you some ideas:
http://machinelearningmastery.com/feature-selecti

laxmi k September 3, 2016 at 2:05 pm #

REPLY

I want it in matlab.

Jason Brownlee September 4, 2016 at 5:19 am #

REPLY

Sorry, I don't have examples in matlab.

Jaggi September 20, 2016 at 5:53 am #

Hello Jason,

As per my understanding, when we speak of 'dimensic
Curse of dimensionality is sort of sin where dimensions
algorithms are not robust enough to handle such high

To reduce the dimension or features, we use algorithm
combination of existing features which try to explain m

Question: Since, these components are created using
complexity is reduced ? How it is beneficially?
Say, there are 10000 features, and each component i.e
Features didn't reduced rather a mathematical combin

Without PCA: GoodBye ~ 1*WorkDone + 1*Meeting + 1*MileStoneCompleted
With PCA: Goodbye ~ PC1
PC1=0.7*WorkDone + 0.2*Meeting +0.4*MileStoneCompleted

Jason Brownlee September 20, 2016 at 8:37 am #

REPLY

Yes Jaggi, features are dimensions.

We are compressing the feature space, and some information (that we think we don't need) is/may be lost.

You do have an interesting point from a linalg perspective, but the ML algorithms are naive in feature space, generally. Deep learning may be different on the other hand, with feature learning. The hidden layers may be doing a PCA-like thing before getting to work.

sai November 13, 2016 at 11:43 pm #

Is there any Scope for pursuing PhD in feature selection?

Jason Brownlee November 14, 2016 at 7:43 am #

REPLY

There may be Sai, I would suggest talking to your advisor.

Poornima December 6, 2016 at 6:29 pm #

REPLY

What would be the best strategy for feature selection in case of text mining or sentiment analysis to be more specific. The size of feature vector is around 2

Jason Brownlee December 7, 2016 at 8:55 a

Sorry Poornima, I don't know. I have not c

Lekan December 22, 2016 at 6:31 am #

How many variables or features can we use i
using Cuckoo Search algorithm on predicting students

Jason Brownlee December 22, 2016 at 6:39 am #

REPLY

There are no limits beyond your hardware or those of your tools.

Arun January 11, 2017 at 2:21 am #

REPLY

can you give some java example code for feature selection using forest optimization algorithm

Jason Brownlee January 11, 2017 at 9:28 am #

REPLY

Sorry Arun, I don't have any Java examples.

Amina February 17, 2017 at 4:07 am #

REPLY

Pls is comprehensive measure feature selection also part of the methods of feature selection?

Jason Brownlee February 17, 2017 at 10:01 am #

REPLY

Hi Amina, I've not heard of "comprehensive measure feature selection" but it sounds like a feature selection method.

Birendra February 28, 2017 at 10:06 pm #

Hi Jason,
I am new to Machine learning. I applied in sklearn RFE
It's giving me error. Is there any way to reduce features

Jason Brownlee March 1, 2017 at 8:37 am #

Yes, this post describes many ways to rec

What is your error exactly? What platform are you

Abdel April 6, 2017 at 6:37 am #

REPLY

Hi Jason,

what is the best method between all this methods in prediction problem ??

is LASSO method great for this type of problem ?

Jason Brownlee April 9, 2017 at 2:37 pm #

REPLY

I would recommend you try a suite of methods and see what works best on your problem.

Al April 26, 2017 at 6:05 pm #

REPLY

Fantastic article Jason, really appreciate this i

If, for example, I have run the below code for feature selection:

test = SelectKBest(score_func=chi2, k=4)
fit = test.fit(X_train, y_train.ravel())

How do I then feed this into my KNN model? Is it simply a case of:

knn = KNeighborsClassifier()
knn.fit(fit) –is this where the feature selection comes in?
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
metric_params=None, n_jobs=1, n_neighbors=5, p=2,
weights='uniform')
predicted = knn.predict(X_test)

Jason Brownlee April 27, 2017 at 8:36 am #

This post may help:
http://machinelearningmastery.com/feature-selecti

Nisha t m May 14, 2017 at 2:21 am #

Sir,
I have multiple data set. I want to perform LASSO regr t
[0,1] vector set as output?

Jason Brownlee May 14, 2017 at 7:31 am #

REPLY

That really depends on your chosen library or platform.

Simone May 30, 2017 at 6:51 pm #

REPLY

Great post!

If I have well understood step n°8, it' s a good procedure *first* applying a linear predictor, and then use a non-linear predictor with the features found before. Is it correct?

Jason Brownlee June 2, 2017 at 12:34 pm #

REPLY

Try linear and nonlinear algorithms on raw
best.

akram June 10, 2017 at 6:03 am #

REPLY

hello Jason Brownlee and thank you for this post,
i'am working on intrusion detection systems IDS, and i want you to advice me about the best features selection algorithm and why?
thanks in advance.

Jason Brownlee June 10, 2017 at 8:30 am #

REPLY

Sorry intrusion detection is not my area of

I would recommend going through the literature an

Yosra Abdelhalim January 28, 2021 at

I would like to build an Intrusion detec
python and the libraries I have to use for ML ;
what I need to learn, any information will be he

Thank in advance

Jason Brownlee January 29, 2021 at 6:03 am #

REPLY

Yes, start here:
https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

And here:
https://machinelearningmastery.com/start-here/#deeplearning

Yosra Abdelhalim January 29, 2021 at 8:25 am #

I really appreciate it, thank you

karthika July 28, 2017 at 6:43 pm #

REPLY

please tell me the evaluation metrics for featu

Jason Brownlee July 29, 2017 at 8:10 am #

REPLY

Ultimately the skill of the model in making predictions.

That is the goal of our project after all!

Swati July 31, 2017 at 4:19 am #

REPLY

Hi!

I have a set of around 3 million features. I want to appl
reduction. How do I do that? I'm using MATLAB.

Swati July 31, 2017 at 4:23 am #

When I use the LASSO function in MATL
corresponding responses) as inputs, I obtain an nx
utilise this output.

Jason Brownlee July 31, 2017 at 8:21 a

Sorry, I cannot help you with the matlab implementations.

Jason Brownlee July 31, 2017 at 8:20 am #

REPLY

Perhaps use an off-the-shelf efficient implementation rather than coding it yourself in matlab?

Perhaps Vowpal Wabbit:
https://github.com/JohnLangford/vowpal_wabbit

Swati July 31, 2017 at 3:20 pm #

REPLY

Thanks

V. Pavithra January 8, 2020 at 4:51 ar

REPLY

Hello

Am doing my PhD in data ming for diseases prediction which features selection is best?

Jason Brownlee January 8, 2020 at 8:33 am #

Good question, this will help:
https://machinelearningmastery.com/faq/single-faq/what-feature-selection-method-should-i-use

Elakkiya September 5, 2017 at 8:45 pm #

HI..
I need your suggestion on something. just assume i ha
f2,f3 set. Each set produce different different output re
feature set. any mathematical way to assign weight to

Jason Brownlee September 7, 2017 at 12:44

Yes, this is what linear machine learning a

Elakkiya September 8, 2017 at 3:39 pm #

REPLY

Thank you. Still its difficult to find how regression algorithm will useful to assign weight .
Can you suggest any material or link to read…

Jason Brownlee September 9, 2017 at 11:51 am #

REPLY

Search linear regression on this blog.

Marie J October 4, 2017 at 5:18 am #

REPLY

Hi Jason! Thank you for your articles – you've been teaching me a lot over the past few weeks. 🙂

Quick question – what is your experience with the best sample size to train the model? I have 290 features and about 500 rows in my dataset. Would this be considered adequate? Or is the rule of thumb to just try and see how well it performs?

Many thanks!

REPLY

Jason Brownlee October 4, 2017 at 5:51 am #

Great question, see this post on the topic:
https://machinelearningmastery.com/much-training-data-required-machine-learning/

REPLY

gene October 18, 2017 at 6:02 pm #

Hello Jason,

I am still confused about your point regarding the featu
the first link you suggested, the advice was to take out
Next start model selection on the remaining data in the

Jason Brownlee October 19, 2017 at 5:34 am

See this post on the difference between t
https://machinelearningmastery.com/difference-tes

Does that help?

REPLY

gene October 22, 2017 at 12:18 am #

Yes, thanks for this post.
But in practice is there any way to integrate feature selection in model selecction while using GridSearchCV in scikit-learn ?

REPLY

Jason Brownlee October 22, 2017 at 5:30 am #

Yes, you could use a Pipeline:
https://machinelearningmastery.com/automate-machine-learning-workflows-pipelines-python-scikit-learn/

Sara May 9, 2019 at 11:05 am #

Hi Jason, thank you, I have le

I would like to integrate feature selection in model selection, as you are saying,"It is important to consider feature selection a part of the model selection process."
I tried to use a scikit-learn Pipeline as you recommended in above. However, pipeline is like a black box, and I cannot follow what it is doing.

I am trying to integrate feature selection (RFECV) as loop inside model selection (gridsearchCV) as below:

```
param_grid = [{'estimator__C': [0.01, 0.1, 1, 10.0, 100, 1000]}]
F1 = RFECV(estimator=svm.SVR(kernel="linear"), step=1)
M1 = GridSearchCV(estimator=F1, param_grid, cv=5)
M1.fit(X_train, y_train)
print(M1.best_params_)
print(M1.best_score_)
print(M1.best_estimator_)
```

This code doesnot give errors, BUT, is
selection?

Jason Brownlee May 9, 2019

Hmmm, too much CV going c

A good pipeline might be [[data prep] -
whole lot.

Sara May 10, 2019 at 6:43 am #

I applied grid search CV to a pipeline, and I get error.

That would be great if you could look at the below error:

```
pipeline1 = Pipeline([ ('feature_selection',SelectFromModel(svm.SVC(kernel='linear'))),
('filter' , SelectKBest(k=11)),
('classification' , svm.SVC(kernel='linear'))
])
gridparams = [{'C': [0.01, 0.1, 1, 10, 100, 1000]}]
model = GridSearchCV(pipeline1, gridparams, cv=5)
model.fit(X, y)
```

However it gives this error:
ValueError: Invalid parameter estimator for estimator Pipeline(memory=None,
steps=[('feature_union', FeatureUnion(n_jobs=None,
transformer_list=[('filter', SelectKBest(k='all', score_func=)), ('feature_selection',
SelectFromModel(estimator=SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
decision_function_shape='ovr…r', ma

shrinking=True, tol=0.001, verbose=False))]). Check the list of available parameters with `estimator.get_params().keys()`.

Jason Brownlee May 10, 2019 at 8:21 am #

Sorry, I don't have the capacity to debug your example.

I have some suggestions here:
https://machinelearningmastery.com/faq/single-faq/can-you-read-review-or-debug-my-code

yerart July 25, 2019 at 5:04 am #

Well Sara, Jason, I did 🙂

Perhaps Sara after all this time has so

First of all, I managed to reproduce the
followed the manual (I mean, Jason's
https://machinelearningmastery.com/a
python-scikit-learn/ and scikit-learn on
when it came to fit … the same error w

So as usual I became curious and afte
learn and Python) … I got it! I cannot h
the error messages. When I did I run e
estimator and found out that the param
show you.

Sara, you're using the same estimator, i.e SVC, for the wrapper feature selection and the classification task on your dataset (by the way it takes ages to fit that) . Both of them have a C hyperparameter. So when you define your param grid and you name 'C' the hyperparameter you want to grid … which C is what you are telling GridSearchCV to iterate? The C from the estimator you use in the wrapper phase or the C in the classification phase of the pipeline? That the same unsolved question GridSearchCV asked itself when fitting and what yields the error.

So what Sara has to do is run model..get_params().keys() and locate the names of the params that end in "__C" and choose the full name of the one she wants and change the name in the param grid definition.

You're welcome 😉

PD: This was my try:

```
1  from sklearn.feature_selection import SelectFromModel
2  from sklearn.svm import SVC
3  from sklearn.model_selection import GridSearchCV
4
5
```

```
#::NOTE::I use feature union for the feature extraction/selection phase
features_sara = []
features_sara.append(('wrapper',
                      SelectFromModel(
                          estimator = SVC(kernel = 'linear', random_state =
                      )
                     )
                    )
#::NOTE::SelctKbest default is f_classif
features_sara.append(
    ('filter', SelectKBest(k = 6))
)
feature_union_sara = FeatureUnion(features_sara)


#::GMG::I build carefully the estimators
estimators_sara = []
estimators_sara.append(
    ('feature_union', feature
)
estimators_sara.append(
    ('classification', SVC(ke
)



#::GMG::Now the pipeline
pipeline_sara = Pipeline(esti


gridparams_sara = {'classifi


model_sara = GridSearchCV(est
                          par
                          cv
                          sco
                          ref
                          return_train_score = False,
                          iid = False
)



#::GMG::And .. I get the SAME error :))
model_sara.fit(X, y)


# **::FIX::** The problem is which C is the `gridparams_sara` defining. If
```


Jason Brownlee July 25, 2019 at 7:59 am #

Impressive, thanks for sharing!

gene November 12, 2017 at 8:32 am #

REPLY

Hello again!
my feature space is over 8000 attributes. When applying RFE, how can I select the right number of feature?
By constructing multiple classfiers (NB, SVM, DT) each of which returns different results.

Jason Brownlee November 12, 2017 at 9:11 am #

REPLY

There is no one best set or no one best model, just many options for you to compare. That is the job of applied ML.

Try building a model with each set and see which is more skillful on unseen data.

gene November 12, 2017 at 7:44 pm #

I want to publish my results. Is it ok to
set with a different number of top features?

Jason Brownlee November 13, 201

When reporting results, you shou
can reproduce your results.

gene November 13, 2017 at 6:44 pm #

REPLY

Yes I understand that, but I meant does that outcome look reasonable?

Hardik Sahi January 8, 2018 at 12:12 pm #

REPLY

Hi
I am getting a bit confused in the section of applying feature selection in cross validation step.
I understand that we should perform feature selection on a different dataset [let's call it FS set] than the dataset we use to train the model [call it train set].

I understand the following steps:
1) perform Feature Selection on FS set.
2) Use above selected features on the training set and fit the desired model like logistic regression model.
3) Now, we want to evaluate the performance of the above fitted model on unseen data [out-of-sample data, hence perform CV]

For each fold in CV phase, we have trainSet and ValidSet. Now we have to again perform feature selection for each fold [& get the features which may/ may not be same as features selected in step 1]. For this, I again have to perform Feature selection on a dataset different from the trainSet and ValidSet.

This is performed for all the k folds and the accuracy is averaged out to get the out-of-sample accuracy for the model predicted in step 2.

I have doubts in regards to how is the out-of-sample accuracy (from CV) an indicator of generalization accuracy of model in step 2. Clearly the feature sets used in both steps will be different.

Also, once I have a model from Step 2 with m<p features selected. How will I test it on completely new data [TestData]? (TestData is having p features and the model is trained on data with m features. What happens to the remaining p-m features??)

Thanks

Jason Brownlee January 8, 2018 at 3:54 pm

A simple approach is to use the training d

I would suggest splitting the training data into train
CV fold (automatically).

Once you pick a final model+procedure, fit on the t
check.

Molood March 9, 2018 at 8:07 am #

Thank you Jason for your article, it was so helpful! I'm working on a set of data which I should to find a business policy among the variables. are any of these methods which you mentioned unsupervised? there's no target label for my dataset. and if there is unsupervised machine learning method, do you know any ready code in github or in any repository for it?

Jason Brownlee March 10, 2018 at 6:13 am #

REPLY

Perhaps a association algorithm:
https://en.wikipedia.org/wiki/Association_rule_learning

Rag March 15, 2018 at 5:24 pm #

REPLY

Sir, Is there any method to find the feature important measures for the neural network?

Jason Brownlee March 16, 2018 at 6:10 am #

REPLY

There may be, I am not across them sorry. Try searching on google scholar.

Yosra March 22, 2018 at 4:35 am #

REPLY

Thank you for the helpful introduction. However, do you have any code using particle swar optmization for features selection ?

Yosra March 22, 2018 at 4:36 am #

Code with R

Jason Brownlee March 22, 2018 at 6:26 am #

Sorry, I do not.

Satya March 28, 2018 at 1:20 am #

Good Morning Jason,
A very nice article. I have a quick question. Please help me out. I am using the R code for Gradient Descent available on internet. The data set ‘women’ is one of the R data sets which has just 15 data points. Here is how I am calling the gradient descent.

n = length(women[,1])
x = women[,1]
y = women[,2]

gradientDesc(x, y, 0.00045, 0.0000001, n, 25000000)

It takes these many iteration to converge and that small learning rate. It is not converging for any higher learning rates. Also I tried using the feature scaling (single feature) as follows but it did not help also – scaling may not be really applicable in case of a single freature

x = (women[,1] – mean(women[,1]))/max(women[,1])

Any help is highly appreciated

Thanks
Satya

Jason Brownlee March 28, 2018 at 6:28 am #

REPLY

Perhaps ask the person who wrote the code about how it works?

Satya March 28, 2018 at 1:22 am #

REPLY

By the way 0.00045 is the learning rate and 0.0000001 is the threshold

Sara April 21, 2018 at 12:57 pm #

Hi,
Is it correct to say that PCA is not only a dimension red
too as in PCA, feature with lower loading should be ex

Jason Brownlee April 22, 2018 at 5:56 am #

Yes, we can treat dimensionality reductio

yerart July 25, 2019 at 8:43 pm #

Jason is right in using “synonym”. As usual the devil is in the details. So I’ve learnt so far. PCA has the small issue of “interpretability”. You got a number of new features (some people would call that feature extraction), ideally much much less than the number of original features. Those new features are a (linear) combination of the original features weighted in a special way. So if you really have (deep) domain knowledge then you can give meaning to those new features and hopefully explain the results the model yields using them. With the feature selection approach (filter, wrapper, embed or a combination thereof) you got a ranked list or a subset of ideally important and non-redundant original features you can explain (even reading the dataset metadata description if available even if you have a shallow domain knowledge) when used in a model.

I’ve seen in meteo datasets (climate/weather) that PCA components make a lot of sense. But what about say genomics? Whoa 🙂

PD: there are ways of make some sense somehow within the “principal components” involving awful things like biplots and loadings that I don’t understand at the moment (and don’t know if I ever will …)

Jason Brownlee July 26, 2019 at 8:22 am #

REPLY

Almost always the features are not interpretable and are best treated as a projection that is there to help the model better learn the structure of the mapping problem.

Sarah June 11, 2018 at 6:39 am #

REPLY

Hi Jason,

Great and useful article – thank you!!

So I've been performing elastic net and gradient boosting machine analyses on my data. Those are methods of feature selection, correct? So, would it be advisable to choose the significant or most influential predictors and include those as the only predictors in a new elast

Sarah June 11, 2018 at 6:57 am #

Also, glmnet is finding far fewer significan
conservative glmnet? Thank you!

Jason Brownlee June 11, 2018 at 1:50 p

Simpler models are preferred in gene
less likely to overfit.

Jason Brownlee June 11, 2018 at 1:49 pm #

REPLY

Generally, I recommend testing a suite of methods on your problem in order to discover what works best.

Sarah Paul June 12, 2018 at 8:59 am #

REPLY

Thank you for your answer!
But, should I use the most influential predictors (as found via glmnet or gbm. etc.) as the only predictors in a new glmnet or gbm (or decision tree, random forest, etc.) model? That doesn't seem to improve accuracy for me. And/or, is it advisable to use them as input in a non-machine learning statistical analysis (e.g., multinomial regression)?
Thank you again!

Jason Brownlee June 12, 2018 at 2:26 pm #

REPLY

Ideally, you only want to use the variables required to make a skilful model.

Try the suggested parameters and compare the skill of a fit model to a model trained on all parameters.

Pratik June 14, 2018 at 4:59 pm #

REPLY

Hi Jason thanks for a wonderful article!!

I need to find the correlation between specific set of fe
correlation of all the features with respect to only class

Jason Brownlee June 15, 2018 at 6:42 am #

You could use the chi-squared independe
https://machinelearningmastery.com/chi-squared-t

Anthony The Koala June 27, 2018 at 8:46 pm #

Dear Dr Jason,
Could you please make the distinction between feature selection (to reduce factors) for predictive modelling and pruning convolutional neural networks (CNN) to improve execution and computation speed please.
Thank you,
Anthony of Sydney

Jason Brownlee June 28, 2018 at 6:18 am #

REPLY

What is “pruning CNNs”?

Anthony The Koala June 28, 2018 at 11:50 am #

REPLY

As I understand, pruning CNNs or pruning convolutional neural networks is a method of reducing the size of a CNN to make the CNN smaller and fast to compute. The idea behind pruning a CNN is to remove nodes which contribute little to the final CNN output. Each node is assigned a weight and ranked.

Those nodes with little weight are eliminated. The result of a smaller CNN is faster computation. However there is a tradeoff in accuracy of matching an actual object to the stored CNN. Software and papers indicate that there is not one method of pruning:

Eg 1 https://www.tensorflow.org/api_docs/python/tf/contrib/model_pruning/Pruning

Eg 2 an implementation in keras, https://www.reddit.com/r/MachineLearning/comments/6vmnp6/p_kerassurgeon_pruning_keras_models_in_python/

Eg 3 a paper https://arxiv.org/abs/1611.06440 it is not the only paper on pruning.

What is the corollary of pruning CNNs and your (this) article? Answer: pruning CNNs involve removing redundant nodes of a CNN while pruning variables in a model as in Weka https://machinelearningmastery.com/feature-se
time/ reduces the number of variables in a mo

Yes they are completely different topics, but th

Thank you,
Anthony of Sydney

Jason Brownlee June 28, 2018 at 2

I see, like classical neural networ

Pruning operates on the learned model, in
on the input to the model.

That is the difference, model and input data.

They are reducing complexity but in different ways, ‘a mapping that has already been learned’ and ‘what could be mapped’ to an output. One fiddles in a small area of hypothesis space for the mapping, the other limits the hypothesis space that is being searched.

Ellen July 17, 2018 at 3:36 am #

REPLY

Hi, Jason. I find your articles really helpful. I have a problem that’s highly related to feature selection, but not the same. So there are correlations between my features which I think if I apply standard feature selection methods, would result in that some features get small weights/importance because they are correlated with a chosen one and thus seem redundant. But I was wondering if you have suggestions for methods that do take into account of feature correlation and assign relatively equal weights/importance to features that are highly correlated?

Jason Brownlee July 17, 2018 at 6:20 am #

Ensembles of decision trees are good at handing irrelevant features, e.g. gradient boosting machines and random forest.

Yanda Jiang April 17, 2020 at 8:06 am #

REPLY

Hi Jason,

I am curious will the feature selection of ensemble learning, like random forest, be done before building tree or each time of node splitting? In another word, I want to know will all of features be used for decision tree during the process or just those selected beforehand?

Many thx

Jason Brownlee April 17, 2020 at 1

Good question.

Ensembles of decision trees, like random f
that the result is an set of trees that only m
to making a prediction – a type of automat
process.

Chris August 7, 2018 at 10:26 am #

Nice write up. What I've found is that the most important features (Boruta and Recursive feature elimination) in my data tend to have the lowest correlation coefficients, and vice versa. Can you shed some light on what's going on?

Jason Brownlee August 7, 2018 at 2:31 pm #

REPLY

Thanks.

It's hard to tell, perhaps a quirk of your dataset?

Harsh August 16, 2018 at 2:51 pm #

REPLY

nice article , really very helpful

you have written inadvertently introduce bias into your models which can result in overfitting. but high bais model mainly underfit the traning data

please correct me if i am worng

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

REPLY

Can you elaborate on what I have “inadvertently written”?

Guru October 18, 2018 at 11:37 pm #

REPLY

Hi Jason, I have one query regarding the below statement

“It is important to consider feature selection a part of th
inadvertently introduce bias into your models which ca

If we have the bias in our model then it should underfit
does bias results in overfitting.

Jason Brownlee October 19, 2018 at 6:05 am

No, a bias can also lead to an overfit. A b
hurtful direction.

Using the test set to train a model as well as the tr
model perform better, but any evaluation on the te
leakage. More here:
https://machinelearningmastery.com/data-leakage-machine-learning/

Yoshitaka November 21, 2018 at 5:38 pm #

REPLY

How do you determine the cut off value when using the feature selection from RandomForest width Scikit-learn and XGBoost’s feature importance methods?

I just choose by heuristic, just feeling.
I thought using grid search or some other optimized methods are better.

Jason Brownlee November 22, 2018 at 6:22 am #

REPLY

Trial and error and go with the cut-off that results in the most skillful model.

cnRZcn December 17, 2018 at 12:28 am #

REPLY

How we can combine the different feature vectors (feature weighting)? If we have two or three different sized feature vectors obtained from our image, how we can combine these features?

Jason Brownlee December 17, 2018 at 6:23 am #

REPLY

You don't, choose one that results in the model with the best performance. I explain more here: https://machinelearningmastery.com/faq/single-faq/what-feature-selection-method-should-i-use

Jasio Zauha December 19, 2018 at 3:03 am #

Hi Jason,

This is a very well written and concise article. What wo
magnitude of effect imposed by changing A to B: shou
B or should I instead provide one array of differences (

Jason Brownlee December 19, 2018 at 6:37

Perhaps try a sensitivity analysis and vary

Win Pa Pa San January 24, 2019 at 3:27 am #

REPLY

Dear Jason;
Thank for explaining about to understand the different between regression and classification.

Jason Brownlee January 24, 2019 at 6:47 am #

REPLY

I explain the difference here:
https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/

Sufyan Danish February 13, 2019 at 1:37 am #

REPLY

hello, sir, I hope u will be in good condition…

kindly guide me that how to use the principal component analysis in weka…
I know how to apply PCA but after applying this I can not know how to use, process, save data and how can I give it to the machine learning algorithm

REPLY

Jason Brownlee February 13, 2019 at 8:01 am #

Sorry, i don't think I have an example of using PCA in Weka.

I do have material on PCA here though:
https://machinelearningmastery.com/calculate-principal-component-analysis-scratch-python/

Sangeeth March 3, 2019 at 5:30 am #

Hi, Thank you for this article. Is Taken's Embe
the input space, a filter approach?. Thanks

Jason Brownlee March 3, 2019 at 8:04 am #

Wonderful question!

No, it is related, but it is probably "feature extractio

REPLY

Avijit March 21, 2019 at 1:07 am #

Sir, can you please discuss something about "Hybrid Feature Selection (HFS-SVM)" with an example? It will be a great help.

REPLY

Jason Brownlee March 21, 2019 at 8:17 am #

Thanks for the suggestion.

What is Hybrid Feature Selection (HFS-SVM) exactly?

REPLY

Avijit March 22, 2019 at 1:05 am #

Mix of multiple feature selection technique .

Jason Brownlee March 22, 2019 at 8:30 am #

REPLY

Okay, thanks.

Avijit March 22, 2019 at 10:46 pm #

Thank you too

Saeed Ahmed March 25, 2019 at 3:49 pm #

Is it possible if we applied feature selection al
every fold, so my question is that can we train the mod

Jason Brownlee March 26, 2019 at 8:00 am #

In that case, you are testing the methodol
approach.

Saeed Ahmed March 26, 2019 at 11:23 am #

If i used the SVM classifier then there is two confusion, first one if we applied Feature selection algorithm at every Fold it may be to select different feature at every Fold then how to find optimized c and g values because the Fold 1 data may be different than Fold 2 and so on. Second one if different features are selected in every fold then if we check the final model on unseen data or independent data then which feature should be selected from independent data.

Jason Brownlee March 26, 2019 at 2:18 pm #

REPLY

Generally, the CV process tests the procedure for selecting features/tuning, rather than a specific set of features/configs – yet you can use it this latter way if you wish by taking the average across the folds.

Shehwar March 27, 2019 at 10:03 pm #

REPLY

Great Tutorial !

I am student of BSCS and trying to discover Keras, Tensorflow. My Question is How can we know which features are selected in training when making KERAS CNN CLASSIFICATION model ?

Jason Brownlee March 28, 2019 at 8:13 am #

REPLY

Great question, the answer is that the selected features result in a better performing model.

You must discover what features result in the best performing model, and what model to use, and what data, etc. This is the challenge of applied machine learning.

Gary April 10, 2019 at 6:55 pm #

Hi,

Is there a recommended way/best practice for querying

Say I create a model with 10 features but then I want t

Is there a model that best suits this use-case ?

I have tried a linear classifier but it needs all 10 feature

You can add real values for the 5 features but a media
values.

Any recommendations ?

Thanks & Rgds,

Gary

Jason Brownlee April 11, 2019 at 6:33 am #

REPLY

Perhaps you can se a model that supports missing values or a mask over missing values?

Perhaps train the model to expect 0 sometimes (e.g. random missing values)? Then provide 0 values for missing values?

Perhaps try training the model with imputed values for missing values, and same as above?

Get creative, try things! Let me know how you go.

Ellie April 18, 2019 at 7:15 am #

REPLY

Hello Jason and Thank you for posting extrem
I have two questions.

If we compare different feature selection methods using a dataset and one of our measures in selecting the best method is how robust the selected feature set is, then Can we do that by testing the model built on an external test set and comparing the training accuracy with the test accuracy to see if we can still gain a good accuracy on the external test set?

If the approach I am taking to measure the robustness of the selected features by feature selection methods is right, then how I can do that for PCA? Should I make the components for all data points including the external dataset? it does not seem right, though.

Your help is appreciated,
Ellie

Jason Brownlee April 18, 2019 at 8:57 am #

Yes, I think model performance is the only
methods.

More here:
https://machinelearningmastery.com/faq/single-faq

ZH May 3, 2019 at 3:17 am #

Hi Jeson, thanks for this great article!
I have confusion where you say in this article:
“A mistake would be to perform feature selection first t
and training on the selected features.”
I believed that performing feature selection first and then perform model selection and training on the selected features, is called filter-based method for feature selection.
– So why this is a mistake?
– Is this a mistake to use Filter-based method which relies only on data set and is classifier-independent?

-“Including feature selection within the inner-loop when using cross-validation&grid-search for model selection”, means that we do feature selection while creating model, is this called embedded method?

Jason Brownlee May 3, 2019 at 6:22 am #

REPLY

It would lead to data leakage:
https://machinelearningmastery.com/data-leakage-machine-learning/

yerart July 25, 2019 at 5:56 am #

REPLY

Jason, I've read your post on data leakage. And that's OK. I think I understand the concept and the need for using pipelines to avoid them. But I think somehow ZH's question still stands for me. Excuse me if this is a silly question but I'm a beginner here.

As I understand a filter approach to feature selection is model neutral downstream the workflow. So say, framing the context, if I want to use a chi2, f_classif or mutual information feature selection (filter or uni-variate as they called it in scikit learn) as a prep data step … why should I put it within a pipeline that then is going to be cross validated for model selection or hyperparameter optimization as good pratice and not doing it "independently" beforehand? Is it related to the "knowledge" I apply on the feature selection phase somehow being leaked to the model selection/hyperparameter optimization phase?

And a related question about the dataset cleaning phase, where you detect and remove or imput NAs and outliers. Say, I have a dataset like the
posts. It has NAs or outliers depending on the
remove and impute the outliers as prep data p
phase produce data leakage? Or is it OK to do
doing the machine learning prep (feature selec
whatnot) proper?

Jason Brownlee July 25, 2019 at 8

Awesome discussion.

Yes, using all data to select features gives

Features selection within the fold contains
and model fitting, not just model fitting.

Ideally all model-based data prep procedures would occur in this way. Finding NAs does not fall under this as it is knowledge-less. Imputing with a mean would require using a mean calculated on the training set within the fold though.

yerart July 27, 2019 at 11:38 pm #

OK brilliant! Yeah, imputation is a potential leakage. And come to think of it, what if the data cleaning task consists of removing the samples with the outliers, not imputing values? I'm thinking of the pima indians database that have some features with outliers. Some of the features outliers (zeros) may be removed because they amount to few samples. It's worth noting that the effect of the removal on the (target) neg/pos (diabetes) subsamples is different (in number). Does this operation on the whole data done before split leak?

Jason Brownlee July 28, 2019 at 6:47 am #

Technically yes.

You want a procedure/knowledge that only operates from the training set. The cross validation tests the procedure of data prep + fitting.

In some cases, the knowledge might be general to the domain – e.g. 0 in this column always means …

Pranav Kamat September 8, 2019 at 2:37 am

Hi Jason,
I am a beginner in ML.
Currently I am working on a regression problem. It

Which technique should I use for feature selection
I googled and kaggled , broke my head over it but

Btw I have used label encoding on categorical vari

Jason Brownlee September 8, 2019 at 5

Good question.

A chi-squared test is a good start.
https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

I think scikit-learn has support for it.

Pranav Kamat September 8, 2019 at 12:10 pm #

REPLY

Jason ,as far as I have read, chi squared test can be used between a categorical predictor and a categorical target.

But in regression , I have categorical predictor but a continuous target.

Here is where I am in doubt of applying chi square test😟

Please bear with with me as I am a newbie🙏

Thanks in advance

Jason Brownlee September 9, 2019 at 5:11 am #

Good question, I'm not sure off hand, perhaps some research and experimentation is required.

Let me know how you go.

Cathy July 13, 2019 at 1:39 am #

REPLY

Hi, I implemented autoencoder to my project and the AUC increased by 1%. The other performance matrixes also increased a little bit. However, there is a
independent dataset using an autoencoder-based svm
feature. I don't know where things go wrong. Any poss

Jason Brownlee July 13, 2019 at 6:58 am #

Not off hand, you may need to debug the

Bushra July 27, 2019 at 4:16 am #

good evening Dr.

I want ask how can use Machine learning in encrypt plain text

I need steps for implement that, please
and what the Machine Learning will add more than encryption algorithms.

I will wait your answer with great passion

Jason Brownlee July 27, 2019 at 6:13 am #

REPLY

I don't know, sorry. Crypto is not my area of expertise.

Leo October 12, 2019 at 12:38 am #

REPLY

Thanks for the article Jason.
From my understanding, correct me if I'm wrong, wrapper methods are heuristic. Is it then safe to say that they are not optimal since they do not test all the combinations in the powerset of the features?

Jason Brownlee October 12, 2019 at 7:06 am #

REPLY

No, “optimal” is not tractable in practice.

In all cases we are doing a heuristic search (guided search, not enumerating all cases) for a subset of features that result in good model skill.

Also, feature subsets interacts with the model, therefore the search problem is way bigger than we might first think:
https://machinelearningmastery.com/applied-machine-learning-as-a-search-problem/

Deepthi December 9, 2019 at 6:50 pm #

Hi Jason,
I have a doubt, do I need train the data on classificatio
methods, can you clarify me on this

Jason Brownlee December 10, 2019 at 7:29

What do you mean by selecting features

Deepthi December 10, 2019 at 4:42 pm #

Selecting features using L1 regularization or tree models

Jason Brownlee December 11, 2019 at 6:48 am #

REPLY

What do you mean exactly?

Deepthi December 11, 2019 at 4:07 pm #

REPLY

Suppose I have 100 features in my dataset and after statistical pre-processing (fill na,remove constant and low variant features) , we have to select the most relevant features for building models(feature reduction and selection). If I use DecisionTreeclassifier/Lasso regression to select best features , Do I need to train the DecisionTree model /Lasso with the selected features?

Jason Brownlee December 12, 2019 at 6:08 am #

REPLY

Yes.

It is best to test different subests of “good” features to find the subset that works the best with your chosen model.

Sarang December 11, 2019 at 6:50 am #

REPLY

Hi Jason,

I’m creating a prediction model which involves cast of r
movie.

Upon doing so, even a data set as small as 2000 data
right way to reduce them ?

When I try to fit PCA, it still shows approx 1500 compo

Jason Brownlee December 11, 2019 at 7:03 a

Perhaps try an SVD:
https://machinelearningmastery.com/singular-value

Hamed January 2, 2020 at 3:25 pm #

REPLY

Hi Jason,

I have a dataset with 10 features.
Rage of 6 of them is between 1 to 10,0, and 4 of them are between 2500 to 52000.
I have tried to do feature selection, but my results are different when I use normalization before feature selection than feature selection without normalization.
I have also read several tutorials, but none of them clearly mentioned that I need normalization before feature selection or not.
For example, in the following tutorial, the feature ranges are very different, but the author didn’t use normalization.
https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

In my point of view, I think in my case I should use normalization before feature selection; I would be so thankful if you could let me know what your thought is?

Thanks

Jason Brownlee January 3, 2020 at 7:14 am #

REPLY

Try it and if it results in a more skillful model, use it.

Hamed January 3, 2020 at 4:42 pm #

REPLY

What's your suggestion.
In my case Normalization before feature selection or not

Jason Brownlee January 4, 2020 a

My best advice is to use controlle
the approach that results in the most skillfu

Tiago Beltrão Lacerda February 29, 2020 at 11:

Hi Jason,

Thanks for you Blog and Books, they are helping me a

Could you give me some help, please?

I'm working on a dataset with mixed data(categorical a

The categorical data: I transformed into dummies variables.
The numerical data: I applied standardization.

Next, I tried RFE. At a first look every thing is fine but one thing worries me: The fact that not a single numerical feature was chosen.

I don't now if it is real of I did something wrong.

Maybe I have to perform feature selection on Categorical and numerical features separately and then blend the result in some way?

Could you give me some light?

thanks.

Tiago Beltrão Lacerda March 1, 2020 at 12:24 am #

REPLY

Complementing:

I performed I loop(from 1 to number_of_feature) with RFE to find the optimal number of features. It was found that 42 features were that optimum value.

Then, listing all these 42 features, I found that all of them where categorical.

I have 329 categorical features and 28 numerical features and 2456 samples.

Given that proportion(11:1), I was inspecting that most of selected features from RFE was going to be categorical. But how can I be sure that this is correct?

thanks.

Jason Brownlee March 1, 2020 at 5:25 a

Maybe. Follow the results/data.

Jason Brownlee March 1, 2020 at 5:20 am #

You're welcome.

Perhaps explore using feature importance scores f

Perhaps explore using statistical methods for featu
https://machinelearningmastery.com/feature-select

Tiago Beltrão Lacerda March 1, 2020 at 1:50 am #

REPLY

Note: as I was researching in this subject, I found that there are two classes of feature selection algorithms: the minimal-optimal and the all-relevant problem. To my particular problem, I find useful to know the all-relevant features.

I Find that the Boruta algorithm implements this, and the the results seems good so far.

He selected 53 features out of 357, both categorical and numerical that a domain expert agreed in their relevance.

Jason Brownlee March 1, 2020 at 5:25 am #

REPLY

Nice work.

Amir Boutaghou March 1, 2020 at 4:27 pm #

first of all thank you so much for this great article. please i have the following question for you :

when i drop feature that is irrelevent to the problem that i try to solve is this step are called “feature extraction” for example i worked before in project in recommendation system based on rating i had review.csv dataframe with these 4 features (user_id,item_id,rating,comment_review). Because i wanted to create an algorithms (example collaborative filtering) based on rating i don’t need the 4th “comment_review” features since my project is not NLP project so i drop it(comment_review). so is what i just did are considered as features selection(or also called feature elimination)

Notice: as you said i know features selections is process to select subset of features that our model will use .

Jason Brownlee March 2, 2020 at 6:15 am #

No, that is feature selection.

Amir Boutaghou March 2, 2020 at 10:44

sorry i didnt understand your answer. you are answer is “No” but after you said to me feature selection !!

please would you reconfirm me if feature selec we call this step.

Thank in advance for you’r answer and time

Jason Brownlee March 2, 2020 at 1:16 pm #

REPLY

Sorry.

You said you dropped a feature/column and asked if this is feature selection. I said no.

Zineb March 5, 2020 at 11:57 pm #

REPLY

Hi Jason,

I miss your articles, I m a little bit buzy with my Phd.
Why when we process features selection using different models and techniques, we may obtain different result even though we ‘re analyzing the same dataset (same features)? the features are describing the data after all, and not related to the models or techniques, so why we dont obtain the same ranking whatever the technique ?

Thanks a lot.

Jason Brownlee March 6, 2020 at 5:35 am #

REPLY

Different algorithms use the provided data/features in different ways, leading to different results.

henry April 23, 2020 at 12:23 pm #

REPLY

Please what feature selection technique do you recommend for 3D facial expression recognition

Jason Brownlee April 23, 2020 at 1:35 pm #

I don't know off the cuff, perhaps review t

David May 9, 2020 at 1:08 am #

Hi Dr,
Feature Selection should be done before or after oneH
create more features? (if we make some sort of feature
do not belong to the original set I do not know if is ok t

Another question is, it is ok after oneHotEncoder to sc
resultants columns?

Thanks in advance!

Jason Brownlee May 9, 2020 at 6:17 am #

REPLY

Yes, feature selection on raw data prior to encoding transforms.

No need to scale encoded variables. No harm though if you want to be lazy with the implementation.

David May 9, 2020 at 8:52 pm #

REPLY

Thanks for the reply!

But the response leads me to another question. If I do not oneHotEncoding the none-numeric(like Strings) features I couldn't apply some Machine Learning strategies for feature selection (like selectKbest for example). What do you suggest to do?
Does this mean that this type of feature should not be

Jason Brownlee May 10, 2020 at 6:10 am #

REPLY

Labels are ordinal encoded or one hot encoded and feature selection is performed prior to encoding typically, or on the ordinal encoding.

Free string data is encoded using a bag of words or embedding representation. Often feature selection here is more expert driven based on the vocab of words you want to support in the domain, such as a subset of the most used words or similar.

David May 10, 2020 at 7:21 pm #

Many thanks for the response! I think
Sorry, I think I was not very clear in the previou
What I mean was, if I have both categorical an
them I can not apply some feature selection m
I do in that case? Should I apply feature select

Sorry to bother you, and again thanks for the r

Jason Brownlee May 11, 2020 at 5

You're welcome.

Ah, I see. Perhaps encode the variables, then apply feature selection. Compare results to using all features.

Masy May 17, 2020 at 10:28 pm #

REPLY

Hello Jason,

Thanks a lot for your efforts.

I have been in debate with my colleague about feature selection methods and what suits text data most, where he believes that unsupervised methods are better than supervised when tackling textual prediction problems. I have tried a few methods and found a statistical method (chi2) to be the best for my problem, leading to optimal performance. What do you think? Any recommendations, please?

Jason Brownlee May 18, 2020 at 6:15 am #

REPLY

I think you must test a suite of methods and discover what works best for a given dataset rather than guessing about generalities.

Huda May 23, 2020 at 4:29 am #

REPLY

Hi Jason, I am currently experimenting on Feature Selection methods for a dataset. It has 2000×2000 dimension (approximately). I am currently contemplating on whether to use Python or Matlab for selecting features (using methods like PSO, GA and so on). Can you suggest which tools are better? And why. Thanks

Jason Brownlee May 23, 2020 at 6:32 am #

I recommend testing a suite of techniques

Shrooq July 5, 2020 at 1:43 am #

I am working on naive bayes model but i am c

Jason Brownlee July 5, 2020 at 7:06 am #

Perhaps evaluate the model with and without it and compare the performance.

SWAROOP July 10, 2020 at 4:04 pm #

REPLY

Am a beginner in field of ML. I'm confused a little. please help me out of this.
Should we train-test-split, feature select(on training set only) and then train the model or feature select on the whole dataset, train-test-split, and then train the model?

Jason Brownlee July 11, 2020 at 6:00 am #

REPLY

Good question, this will help:
https://machinelearningmastery.com/data-preparation-without-data-leakage/

DataCamp Copy? July 24, 2020 at 12:51 am #

Did you also write the DataCamp tutorial on this topic or give permission for them to copy? It's pretty much a word-for-word copy of this post (with some alterations that actually make it harder to understand/less well-written). Linked here: https://www.datacamp.com/community/tutorials/feature-selection-python

Jason Brownlee July 24, 2020 at 6:30 am #

REPLY

It is a copy without permission.

Some people have no shame.

Yaso Ali October 15, 2020 at 3:23 am #

Thank you for the valuable information;

I'm beginner in this field; So there is question in my mi

Jason Brownlee October 15, 2020 at 6:15 am

Depends on your dataset I guess.

Younes January 11, 2021 at 5:56 am #

REPLY

Could you please introduce me,if there is any machine learning model such as Multivariated Adaptive Regression Spline (MARS) which has an ability to select a few number of predictive variables (when the first data set is huge) by it's interior algorithm?

In other world is there any machine learning model which has an ability to feature selection by it's algorithm (by itself)?

Jason Brownlee January 11, 2021 at 6:23 am #

REPLY

Yes, many linear models offer regularization that perform automatic feature selection (e.g. LASSO).

Also ensembles of decision trees can also perform auto feature selection (e.g. random forest, xgboost).

Younes January 11, 2021 at 6:34 am #

Thank you so much for your reply, please let me know what is your opinion about Partial least Square regression (PLSR)?

Does PLSR select just a number of predict variables and use them for modeling processing?

Jason Brownlee January 11, 2021 at 7:58 am #

REPLY

Sorry, I don't have a tutorial on the topic, perhaps this will help:
https://en.wikipedia.org/wiki/Partial_least_squares_regression

Emmily February 22, 2021 at 4:20 am #

How do you understand and explain the proce
learning algorithm–> performance, by applying concep
vision environment? Explain with an example or any ar

Jason Brownlee February 22, 2021 at 5:05 a

Sounds like a homework or interview que

Emmily February 22, 2021 at 5:09 am #

If you'll solve it, I'll very thankful to you.

Leave a Reply

Name (required)

Email (will not be published) (required)

SUBMIT COMMENT

Welcome!
I'm *Jason Brownlee* PhD
and I **help developers** get results with **machine learning**.
Read more

Never miss a tutorial:

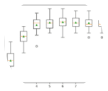
Picked for you:


How to Choose a Feature Selection Method For Machine Learning


Data Preparation for Machine Learning (7-Day Mini-Course)


How to Calculate Feature Importance With Python

Recursive Feature Elimination (RFE) for Feature Selection in P



How to Remove Outliers for Machine Learning

Loving the Tutorials?

The Data Preparation EBook is
where you'll find the ***Really Good*** stuff.

>> SEE WHA