

CSDN

博客 (//blog.csdn.net/)

学院 (//edu.csdn.net/)

下载 (http://download.csdn.net/)

GitChat (http://gitbook.cn/?ref=csdn)

论坛 (http://bbs.csdn.net)

...

写博客

发Chat

登录 (https://passport.csdn.net/account/login)

注册 (https://passport.csdn.net/account/mobileRegister?action=mobileRegister)

特征选择--scikit-learn

原创

2016年07月29日 12:28:12

标签: 机器学习 (http://so.csdn.net/so/search/s.do?q=机器学习&t=blog)

10891

特征选择 ( Feature Selection ) :choosing a subset of all the features(the ones more informative)。最终得到的特征选是原来特征的一个子集。

特征选取是机器学习领域非常重要的一个方向。  
主要有两个功能：

- ( 1 ) 减少特征数量、降维，使模型泛化能力更强，减少过拟合
- ( 2 ) 增强特征和特征值之间的理解

## 1, 去掉取值变化小的特征 ( Removing features with low variance )

这应该是最简单的特征选择方法了：假设某特征的特征值只有0和1，并且在所有输入样本中，95%的实例的该特征取值都是1，那就可以认为这个特征作用不大。如果100%都是1，那这个特征就没意义了。当特征值都是离散型变量的时候这种方法才能用，如果是连续型变量，就需要将连续变量离散化之后才能用，而且实际当中，一般不太会有95%以上都取某个值的特征存在，所以这种方法虽然简单但是不太好用。可以把它作为特征选择的预处理，先去掉那些取值变化小的特征，然后再从接下来提到的的特征选择方法中选择合适的进行进一步的特征选择。

```
1 sklearn.feature_selection.VarianceThreshold(threshold=0.0)
```

```
1 from sklearn.feature_selection import VarianceThreshold
2 X = [[0, 0, 1], [0, 1, 0], [1, 0, 0], [0, 1, 1], [0, 1, 0], [0, 1, 1]]
3 sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
4 sel.fit_transform(X)
5 # array([[0, 1],
6 #        [1, 0],
7 #        [0, 0],
8 #        [1, 1],
9 #        [1, 0],
10 #        [1, 1]])
```

第一列的特征被去掉

## 2, 单变量特征选择 ( Univariate feature selection )

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

面向未来的历史 (http://...)

+关注

(http://blog.csdn.net/a1368783069)

原创

粉丝

喜欢

未开通

67

17

16

(https://gite

少儿编程

5 GB

他的最新文章  
更多文章 (http://blog.csdn.net/a1368783069)

python 处理请求获取的图片 (http://blog.csdn.net/a1368783069/article/details/79093877)

go安装 (http://blog.csdn.net/a1368783069/article/details/78990815)

python Pexpect 实现输密码 scp 拷贝 (http://blog.csdn.net/a1368783069/article/details/78721796)

unicorn 部署 flask 应用 (http://blog.csdn.net/a1368783069/article/details/78665531)

boost安装 c++ (http://blog.csdn.net/a1368783069/article/details/78522405)

### 文章分类

python (http://blog.csdn.net/a...	38篇
c# (http://blog.csdn.net/a1368...	1篇
Video&Audio (http://blog.csdn...	1篇
windows (http://blog.csdn.net/...	1篇
Prezi (http://blog.csdn.net/a13...	3篇

基于单变量统计测试。

展开

单变量特征选择能够对每一个特征进行测试，衡量该特征和响应变量之间的关系，根据得分扔掉不好的特征。对于回归和分类问题可以采用卡方检验等方式对特征进行测试。

方法简单，易于运行，易于理解，通常对于理解数据有较好的效果（但对特征优化、提高泛化能力来说不一定有效）；这种方法有许多改进的版本、变种。

因此建议作为特征选择的前处理中的一步。

```
sklearn.feature_selection.SelectKBest(score_func=<function f_classif>, k=10)
```

选择前k个分数较高的特征，去掉其他的特征。

```
sklearn.feature_selection.SelectPercentile(score_func=<function f_classif>, percentile=10)
```

f\_regression（单因素线性回归试验）用作回归  
chi2\_2方检验，f\_classif（方差分析的F值）等用作分类

```
from sklearn.datasets import load_iris
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
iris = load_iris()
X, y = iris.data, iris.target
X.shape
# (150, 4)
X_new = SelectKBest(chi2, k=2).fit_transform(X, y)
X_new.shape
# (150, 2)
```

选择一定百分比的最高的评分的特征。

```
sklearn.feature_selection.SelectFpr(score_func=<function f_classif>, alpha=0.05)
```

根据配置的参选搜索

```
sklearn.feature_selection.GenericUnivariateSelect(score_func=<function f_classif>, mode='percentile', param=1e-05)
```

### 3,递归特征消除Recursive feature elimination（RFE）

递归特征消除的主要思想是反复的构建模型（如SVM或者回归模型）然后选出最好的（或者最差的）的特征（可以根据系数来选），把选出来的特征选择出来，然后在剩余的特征上重复这个过程，直到所有特征都遍历了。这个过程中特征被消除的次序就是特征的排序。因此，这是一种寻找最优特征子集的贪心算法。

RFE的稳定性很大程度上取决于在迭代的时候底层用哪种模型。例如，假如RFE采用的普通的回归，没有经过正则化的回归是不稳定的，那么RFE就是不稳定的；假如采用的是Ridge，而用Ridge正则化的回归是稳定的，那么RFE就是稳定的。

```
class sklearn.feature_selection.RFECV(estimator, step=1, cv=None, scoring=None, estimator_params=None, verbose=0)
```

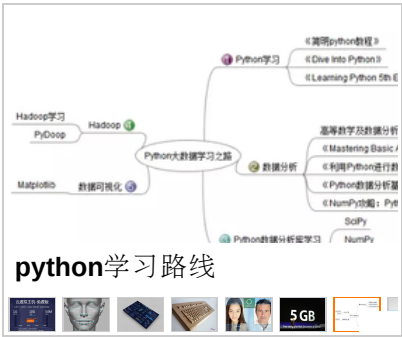
#### 文章存档

2018年1月	(http://blog.csdn.n...	2篇
2017年12月	(http://blog.csdn....	1篇
2017年11月	(http://blog.csdn....	2篇
2017年10月	(http://blog.csdn....	2篇
2017年9月	(http://blog.csdn.n...	1篇

展开

#### 他的热门文章

- python网络爬虫抓取动态网页并将数据存入数据库MySQL (http://blog.csdn.net/a1368783069/article/details/48375695)  
13505
- MiKTeX与Texmaker 配置使用 (http://blog.csdn.net/a1368783069/article/details/46692803)  
11071
- 特征选择--scikit-learn (http://blog.csdn.net/a1368783069/article/details/52048349)  
10831
- python字典多键值及重复键值的使用 (http://blog.csdn.net/a1368783069/article/details/46891685)  
10456
- Graphviz -图形可视化，python实现 (http://blog.csdn.net/a1368783069/article/details/52067404)  
7386



#### 联系我们

网站客服 (http://wpa.qq.com/msgrd?v=3&uin=2431299880&site=qq&menu=yes)

vebmaster@csdn.net (mailto:webmaster@csdn.net)

微博客服 (http://e.weibo.com/csdnsupport/profile)

100-660-0108

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录 注册

关于 招聘 广告服务 阿里云

©2018 CSDN 京ICP证09002463号

(http://www.miibeian.gov.cn/)



经营性网站备案信息

(http://www.hd315.gov.cn/beian

/view.asp?bianhao=010202001032100010)



网络110报警服务 (http://www.cyberpolice.cn/)

```
1 from sklearn.svm import SVC
2 from sklearn.datasets import load_digits
3 from sklearn.feature_selection import RFE
4 import matplotlib.pyplot as plt
5
6 # Load the digits dataset
7 digits = load_digits()
8 X = digits.images.reshape((len(digits.images), -1))
9 y = digits.target
10
11 # Create the RFE object and rank each pixel
12 svc = SVC(kernel="linear", C=1)
13 rfe = RFE(estimator=svc, n_features_to_select=1, step=1)
14 rfe.fit(X, y)
15 ranking = rfe.ranking_.reshape(digits.images[0].shape)
16
17 # Plot pixel ranking
18 plt.matshow(ranking)
19 plt.colorbar()
20 plt.title("Ranking of pixels with RFE")
21 plt.show()
```



## 4. Feature selection using SelectFromModel

SelectFromModel 是一个 meta-transformer，可以和在训练完后有一个coef\_ 或者 feature\_importances\_ 属性的评估器（机器学习算法）一起使用。

如果相应的coef\_ 或者feature\_importances\_ 的值小于设置的阈值参数，这些特征可以视为不重要或者删除。除了指定阈值参数外，也可以通过设置一个字符串参数，使用内置的启发式搜索找到最优阈值。可以使用的字符串参数包括：“mean”，“median” 以及这两的浮点乘积，例如“0.1\*mean”。

```
1 sklearn.feature_selection.SelectFromModel(estimator, threshold=None, pfit=False)
```

与Lasso一起使用，从boston数据集中选择最好的两组特征值。

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 from sklearn.datasets import load_boston
5 from sklearn.feature_selection import SelectFromModel
6 from sklearn.linear_model import LassoCV
7
8 # Load the boston dataset.
9 boston = load_boston()
10 X, y = boston['data'], boston['target']
11
12 # We use the base estimator LassoCV since the L1 norm promotes sparsity of features.
13 clf = LassoCV()
14
15 # Set a minimum threshold of 0.25
16 sfm = SelectFromModel(clf, threshold=0.25)
17 sfm.fit(X, y)
18 n_features = sfm.transform(X).shape[1]
19
20 # Reset the threshold till the number of features equals two.
21 # Note that the attribute can be set directly instead of repeatedly
22 # fitting the metatransformer.
23 while n_features > 2:
24     sfm.threshold += 0.1
25     X_transform = sfm.transform(X)
26     n_features = X_transform.shape[1]
27
28 # Plot the selected two features from X.
29 plt.title(
30     "Features selected from Boston using SelectFromModel with "
31     "threshold %0.3f." % sfm.threshold)
32 feature1 = X_transform[:, 0]
33 feature2 = X_transform[:, 1]
34 plt.plot(feature1, feature2, 'r.')
35 plt.xlabel("Feature number 1")
36 plt.ylabel("Feature number 2")
37 plt.ylim((np.min(feature2), np.max(feature2)))
38 plt.show()
```

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

注册



## 4.1,L1-based feature selection

L1正则化将系数 $w$ 的l1范数作为惩罚项加到损失函数上，由于正则项非零，这就迫使那些弱的特征所对应的系数变成0。因此L1正则化往往会使学到的模型很稀疏（系数 $w$ 经常为0），这个特性使得L1正则化成为一种很好的特征选择方法。

```
1 from sklearn.svm import LinearSVC
3 from sklearn.datasets import load_iris
2 from sklearn.feature_selection import SelectFromModel
iris = load_iris()
4 X, y = iris.data, iris.target
X.shape
# (150, 4)
5 lsvc = LinearSVC(C=0.01, penalty="l1", dual=False).fit(X, y)
6 model = SelectFromModel(lsvc, prefit=True)
X_new = model.transform(X)
X_new.shape
# (150, 3)
```

## 4.2, 随机稀疏模型Randomized sparse models

一些相互关联的特征是基于L1的稀疏模型的限制，因为模型只选择其中一个特征。为了减少这个问题，可以使用随机特征选择方法，通过打乱设计的矩阵或者子采样的数据并，多次重新估算稀疏模型，并且统计有多少次一个特定的回归量是被选中。

RandomizedLasso使用Lasso实现回归设置

```
1 sklearn.linear_model.RandomizedLasso(alpha='aic', scaling=0.5, sample_fraction=0.75, n_resampling=200, selection_thre:
```

RandomizedLogisticRegression 使用逻辑回归 logistic regression，适合分类任务

```
1 sklearn.linear_model.RandomizedLogisticRegression(C=1, scaling=0.5, sample_fraction=0.75, n_resampling=200, selector
```

## 4.3, 基于树的特征选择Tree-based feature selection

基于树的评估器 (查看sklearn.tree 模块以及在sklearn.ensemble模块中的树的森林) 可以被用来计算特征的重要性，根据特征的重要性去掉无关紧要的特征 (当配合sklearn.feature\_selection.SelectFromModel meta-transformer):

```
1 from sklearn.ensemble import ExtraTreesClassifier
2 from sklearn.datasets import load_iris
3 from sklearn.feature_selection import SelectFromModel
iris = load_iris()
4 X, y = iris.data, iris.target
X.shape
5 # (150, 4)
clf = ExtraTreesClassifier()
6 clf = clf.fit(X, y)
7 clf.feature_importances_
array([ 0.04..., 0.05..., 0.4..., 0.4...])
8 model = SelectFromModel(clf, prefit=True)
9 X_new = model.transform(X)
X_new.shape
10 # (150, 2)
```

## 5, Feature selection as part of a pipeline

在进行学习之前，特征选择通常被用作预处理步骤。在scikit-learn中推荐使用的处理的方法是sklearn.pipeline.Pipeline

```
1 sklearn.pipeline.Pipeline(steps)
```

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

注册



Pipeline of transforms with a final estimator.

Sequentially 应用一个包含 transforms and a final estimator的列表，pipeline中间的步骤必须是‘transforms’，也就是它们必须完成fit 以及transform 方法s. final estimator 仅仅只需要完成 fit方法.

使用pipeline是未来组合多个可以在设置不同参数时进行一起交叉验证的步骤。因此，它允许设置不同步骤中的参数使用参数名，这些参数名使用‘\_’进行分隔。如下实例中所示：

```
1 from sklearn import svm
3 from sklearn.datasets import samples_generator
2 from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_regression
4 from sklearn.pipeline import Pipeline
# generate some data to play with
5 X, y = samples_generator.make_classification(
...     n_informative=5, n_redundant=0, random_state=42)
6 # ANOVA SVM-C
anova_filter = SelectKBest(f_regression, k=5)
7 clf = svm.SVC(kernel='linear')
8 anova_svm = Pipeline([('anova', anova_filter), ('svc', clf)])
# You can set the parameters using the names issued
# For instance, fit using a k of 10 in the SelectKBest
9 # and a parameter 'C' of the svm
10 anova_svm.set_params(anova__k=10, svc__C=.1).fit(X, y)
...
12 Pipeline(steps=[...])
prediction = anova_svm.predict(X)
13 anova_svm.score(X, y)
14 0.77...
15 # getting the selected features chosen by anova_filter
anova_svm.named_steps['anova'].get_support()
16 # array([ True,  True,  True, False, False,  True, False,  True,  True, True,
17        False, False,  True, False,  True, False, False, False, False,
18        True], dtype=bool)
```

简单语法示例：

```
1 clf = Pipeline([
2     ('feature_selection', SelectFromModel(LinearSVC(penalty="l1"))),
3     ('classification', RandomForestClassifier())
4 ])
5 clf.fit(X, y)
```

参考：

常用特征选取算法

<http://www.cnblogs.com/wymlnn/p/4569437.html> (<http://www.cnblogs.com/wymlnn/p/4569437.html>)

Feature selection

[http://scikit-learn.org/stable/modules/feature\\_selection.html](http://scikit-learn.org/stable/modules/feature_selection.html) ([http://scikit-learn.org/stable/modules/feature\\_selection.html](http://scikit-learn.org/stable/modules/feature_selection.html))

注：以后继续补充

版权声明：本文为博主原创文章，未经博主允许不得转载。





## [Sklearn应用5] Feature Selection 特征选择（一） SelectFromModel

此内容在sklearn官网地址：[http://scikit-learn.org/stable/modules/feature\\_selection.html](http://scikit-learn.org/stable/modules/feature_selection.html) sklearn版本：0.18.2 ...

 sscc\_learning 2017年06月29日 23:19  1391

([http://blog.csdn.net/sscc\\_learning/article/details/73929038](http://blog.csdn.net/sscc_learning/article/details/73929038))

### 使用sklearn优雅地进行数据挖掘

 wuzhongdehua1 2016年09月12日 18:16  3141

目录: 1. 使用sklearn进行数据挖掘 1.1 数据挖掘的步骤 1.2 数据初貌 1.3 关键技术 2 并行处理 2.1 整体并行处理 2.2 部分并行处理 ...



(<http://blog.csdn.net/wuzhongdehua1/article/details/52515849>)

### 一个数学公式教你秒懂天下英语



老司机教你一个数学公式秒懂天下英语



## sklearn学习笔记2 Feature\_extraction库

 wateryouyo 2016年12月28日 10:43  1730

1. 将字典格式的数据转换为特征。前提：数据是用字典格式存储的，通过调用DictVectorizer类可将其转换成特征，对于特征值为字符串的变量，自动转换为多个特征变量，类似前面提到的onehot编...

(<http://blog.csdn.net/wateryouyo/article/details/53906426>)

## sklearn feature extraction

 perfectmanman 2015年11月03日 14:40  613

文本特征提取词袋（Bag of Words）表征文本分析是机器学习算法的主要应用领域。但是，文本分析的原始数据无法直接丢给算法，这些原始数据是一组符号，因为大多数算法期望的输入是固定长度的数值特征向量...

(<http://blog.csdn.net/perfectmanman/article/details/49616043>)

## 机器学习中的特征选择

 rui307 2016年04月25日 17:24  8366

首先声明，本人个人观点，仅供交流。本人欠专业人士，并不了解显示实践中的特征工程。特征选择是一个重要的数据预处理过程，获得数据之后要先进行特征选择然后再训练模型。主要作用：1、降维 2、去除不相关特...



(<http://blog.csdn.net/rui307/article/details/51243796>)

### 一个数学公式教你秒懂天下英语



老司机教你一个数学公式秒懂天下英语



## Feature selection using SelectFromModel

 FontThrone 2018年01月15日 15:17  32

SelectFromModel sklearn在Feature selection模块中内置了一个SelectFromModel，该模型可以通过Model本身给出的指标对特征进行选择，其作用与其名字...

(<http://blog.csdn.net/FontThrone/article/details/79064930>)

## sklearn的一些总结

 wang1127248268 2016年11月21日 20:41  4954

使用sklearn进行数据挖掘1.1 数据挖掘的步骤1.2 数据初貌1.3 关键技术并行处理并行处理2.1 整体并行处理2.2 部分并行处理流水线处理自动化调参持久化回顾总结参考资料使用sklearn...

(<http://blog.csdn.net/wang1127248268/article/details/53264041>)


加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

注册



## 利用 **scikit-learn** 进行 **FeatureSelection**


 lming\_08 2014年09月11日 19:45 4166

1. >>> from sklearn.datasets import load\_iris >>> from sklearn.feature\_selection import SelectKBe...

([http://blog.csdn.net/lming\\_08/article/details/39210409](http://blog.csdn.net/lming_08/article/details/39210409))


## 记一次失败的 **kaggle** 比赛（2）：构造新特征、特征筛选

接第一篇：<http://blog.csdn.net/mmc2015/article/details/51095446> 第一篇中提到的主要问题：第一：暴力搜索特征的方式在特征数较多的情况下不可...

 mmc2015 2016年04月08日 13:08 4288


(<http://blog.csdn.net/mmc2015/article/details/51095588>)

## 结合 **Scikit-learn** 介绍几种常用的特征选择方法

参考<http://www.cnblogs.com/hhh5460/p/5186226.html> 未完，占坑  q383700092 2016年12月26日 22:20 2120  
后续目标构成一套成型的自动特征选择的多方案集成输出...

(<http://blog.csdn.net/q383700092/article/details/53889936>)


## python 之 **sklearn**

 liujiandu101 2016年06月13日 09:27 7109

Scikit Learn: 在python中机器学习 Warning 警告：有些没能理解的句子，我以自己的理解意译。翻译自：Scikit Learn: Machine Learning...

(<http://blog.csdn.net/liujiandu101/article/details/51654975>)

## **sklearn** 使用总结

 qingqing7 2017年11月29日 08:53 137

本文主要参考Cer\_ml和Jorocco；sklearn是一个数据挖掘的python库，github地址，该库集成了大量的数据挖掘算法，并对数据做预算处理，对算法进行集成和预测结果进行验证和评...

(<http://blog.csdn.net/qingqing7/article/details/78661298>)

## Python 机器学习库 **SKLearn** 的特征选择

 cheng9981 2017年04月30日 17:10 1879

参考地址：[http://scikit-learn.org/stable/modules/feature\\_selection.html#feature-selection](http://scikit-learn.org/stable/modules/feature_selection.html#feature-selection) sklearn.feature...


(<http://blog.csdn.net/cheng9981/article/details/71023709>)

## 结合 **Scikit-learn** 介绍几种常用的特征选择方法 Bryan\_\_ 2016年06月07日 22:51 21958

特征选择(排序)对于数据科学家、机器学习从业者来说非常重要。好的特征选择能够提升模型的性能，更能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。特征选择主要有两个功能：...

([http://blog.csdn.net/Bryan\\_\\_/article/details/51607215](http://blog.csdn.net/Bryan__/article/details/51607215))


## 特征提升之特征筛选

 cicilover 2017年09月05日 18:23 6238

良好的数据特征组合不需太多，就可以使得模型的性能表现突出。冗余的特征虽然不会影响到模型的性能，但使得CPU的计算做了无用功。比如，PCA主要用于去除多余的线性相关的特征组合，因为这些冗余的特征组合不会...

(<http://blog.csdn.net/cicilover/article/details/77854621>)

## **Sklearn** 中的 **f\_classif** 和 **f\_regression**

 jetFlow 2017年12月24日 13:23 183

这两天在看Sklearn的文档，在feature\_selection一节中遇到俩f值，它们是用来判断模型中特征与因变量的相关性的。刚开始看的时候一头雾水，因为需要数理统计中方差分析的背景，现在在这里简...

(<http://blog.csdn.net/jetFlow/article/details/78884619>)

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录

注册

×

## 特征选择和特征理解



ivysister 2016年05月23日 16:52

作者: Edwin Jarvis 特征选择(排序)对于数据科学家、机器学习从业者来说非常重要。好的特征选择能够提升模型的性能,更能帮助我们理解数据的特点、底层结构,这对进一步改善模型、算法都有着重要作...

(<http://blog.csdn.net/ivysister/article/details/51482917>)

## 几种常用的特征选择方法



LY\_sys629 2016年12月14日 16:33 8514

几种常用的特征选择方法

([http://blog.csdn.net/LY\\_sys629/article/details/53641569](http://blog.csdn.net/LY_sys629/article/details/53641569))

## Scikit-learn: 模型选择Model selection之pipeline和交叉验证

<http://blog.csdn.net/pipisorry/article/details/52250983>选择合适的estimator 通常机器学习最难的一部分是选择合适的estimator, 不同...

pipisorry 2016年08月19日 15:15 6802

(<http://blog.csdn.net/pipisorry/article/details/52250983>)

## sklearn逻辑回归(Logistic Regression,LR)类库使用小结

原文出处: <http://www.07net01.com/2016/11/1706402.html>, 在原文的基础上做了一些修订 sklearn中LogisticRegression的API如下, 官方文档...

sun\_shengyun 2016年12月22日 11:36 16883

([http://blog.csdn.net/sun\\_shengyun/article/details/53811483](http://blog.csdn.net/sun_shengyun/article/details/53811483))