



博客 (<http://blog.csdn.net/?ref=toolbar>) 学院 (<http://edu.csdn.net/?ref=toolbar>)

下载 (<http://download.csdn.net/?ref=toolbar>)

GitChat (<http://gitbook.cn/?ref=csdn>) 更多 ▾



4

## pandas使用get\_dummies进行one-hot编码



原创

2016年10月17日 09:30:41



标签：pandas one-hot (<http://so.csdn.net/so/search/s.do?q=pandas one-hot&t=blog>)

17586



离散特征的编码分为两种情况：

1. 离散特征的取值之间没有大小的意义，比如color：[red,blue],那么就使用one-hot编码
2. 离散特征的取值有大小的意义，比如size:[X,XL,XXL],那么就使用数值的映射{X:1,XL:2,XXL:3}

使用pandas可以很方便的对离散型特征进行one-hot编码

[python]

```
1. import pandas as pd
2. df = pd.DataFrame([
3.     ['green', 'M', 10.1, 'class1'],
4.     ['red', 'L', 13.5, 'class2'],
5.     ['blue', 'XL', 15.3, 'class1']])
6.
7. df.columns = ['color', 'size', 'prize', 'class label']
8.
9. size_mapping = {
10.     'XL': 3,
11.     'L': 2,
12.     'M': 1}
13. df['size'] = df['size'].map(size_mapping)
14.
```



weixin\_3506... ▾

(<http://so.csdn.net/so/search/s.do?q=pandas one-hot&t=blog>)

ref=toolbar\_source=csdnblo



BYR\_jiandong (<http://blo...>)

+ 关注

(<http://blog.csdn.net/lujiandong1>)

码云

未开通

原创

227

粉丝

166

喜欢

3

(<https://github.com/lujiandong1>)

utm\_sourc

### 他的最新文章

更多文章 (<http://blog.csdn.net/lujiandong1>)

NDCG及其实现 (<http://blog.csdn.net/lujiandong1/article/details/77123805>)

keras 设置学习率的办法 (<http://blog.csdn.net/lujiandong1/article/details/76610940>)

tensorflow中对于模型的参数都必须声明为变量 (<http://blog.csdn.net/lujiandong1/article/details/73997600>)

keras2 predict和fit\_generator的坑 (<http://blog.csdn.net/lujiandong1/article/details/73556163>)

```

15. class_mapping = {label:idx for idx,label in enumerate(set(df['class label']))}
16. df['class label'] = df['class label'].map(class_mapping)

```

说明：对于有大小意义的离散特征，直接使用映射就可以了，{'XL':3,'L':2,'M':1}

	color	size	prize	class label
0	green	1	10.1	0
1	red	2	13.5	1
2	blue	3	15.3	0

Using the get\_dummies will create a new column for every unique string in a certain column:使用get\_dummies进行one-hot编码

[python]

```
1. pd.get_dummies(df)
```

	size	prize	class label	color_blue	color_green	color_red
0	1	10.1	0	0	1	0
1	2	13.5	1	0	0	1
2	3	15.3	0	1	0	0



发表你的评论

([http://my.csdn.net/weixin\\_35068028](http://my.csdn.net/weixin_35068028))

二叉树中两个节点的最近公共父节点 (<http://blog.csdn.net/lujiandong1/article/details/71438944>)



卡瓦依钢琴

雅马哈钢琴价格

钢琴二手

住人集装箱

## 他的热门文章

安装sqlalchemy时遇到Error: DLL I  
午，终于搞定了！(解决)  
19348

pandas使用get\_dummies进行one-hot编  
码 (<http://blog.csdn.net/lujiandong1/article/details/52836051>)  
17556

SVM的两个参数 C 和 gamma (<http://blog.csdn.net/lujiandong1/article/details/46386201>)  
14646

tensorflow conv2d的padding解释以及参  
数解释 (<http://blog.csdn.net/lujiandong1/a>)



duguodong189 (/duguodong189) 2017-05-16 21:46

1楼

(/duguodong189)

回复

rticle/details/53728053)

📖 11982

import sys sys.path.append(...) (http://blog.csdn.net/lujiandong1/article/details/47159259)

📖 11358

## pandas get\_dumy (http://blog.csdn.net/eshaliu/article/details/53557989)



另一种常用于统计建模或机器学习的转换方式是：将分类变量（categorical variable）转换为“哑变量矩阵”（dummy matrix）或“指标矩阵”（indicator matrix）。如...



eshaoliu (http://blog.csdn.net/eshaliu) 2016年12月10日 13:53 📖 5526



## pandas使用get\_dummies()进行独热编码 (http://blog.csdn.net/sinat\_32547403/article/det...)



离散特征的编码分为两种情况：1、离散特征的取值之间没有大小的意义，比如color：[red,blue],那么就使用one-hot编码 2、离散特征的取值有大小的意义，比如size:[X,XL,X...



sinat\_32547403 (http://blog.csdn.net/sinat\_32547403) 2017年06月07日 13:48 📖 1625



广告

### 一位超过10年工作经验的程序员总结！

这做了10年web开发的程序员的总结分析给你，也可能是我们正在经历的人生。他的这22条总结希望可以帮到你少走弯路...

(http://www.baidu.com/cb.php?c=lgF\_pyfqHmknj0dP1f0IZ0qnfK9ujYzP1ndPWb10Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YYuhczn1lhnHTdPHwWryRz0AwY5HDdnHfLPWbsPjn0lgF\_5y9YIZ0IQzq-uZR8mLPbUB48ugfEIAqspynElvNBnHqdlAdxTvqdThP-5yF\_UvTkn0KzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqPHDknjR)

## pandas get\_dummies()使用 (http://blog.csdn.net/Edwards\_June/article/details/73716580)



### 全部透明手机



```
In [72]: df = pd.DataFrame({'key': ['b', 'b', 'a', 'c', 'a', 'b'], ...: 'data...
```



Edwards\_June ([http://blog.csdn.net/Edwards\\_June](http://blog.csdn.net/Edwards_June)) 2017年06月25日 14:20 2289

## pandas 下的 one hot encoder 及 pd.get\_dummies() 与 sklearn.preprocessing 下的 OneHo..

sklearn.preprocessing 下除了提供 OneHotEncoder 还提供 LabelEncoder ( 简单地将 categorical labels 转换为不同的数字 ); 1. 简单区...



lanchunhui (<http://blog.csdn.net/lanchunhui>) 2017年06月05日 19:17 2595



## pandas中的几个重要函数 (<http://blog.csdn.net/u012969412/article/details/69526183>)

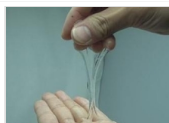
1. read\_csv读取csv文件 import pandas as pd data = pd.read\_csv("./data/a.csv",encoding="gbk")2. get\_dummi...



u012969412 (<http://blog.csdn.net/u012969412>) 2017年04月07日 18:21 877



润滑剂怎么使



性的润滑剂



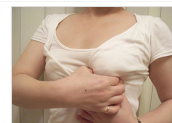
怎样快速受孕



it培训机构排名



botox瘦腿



发育早



达内可靠吗

## python中pandas库学习笔记 (<http://blog.csdn.net/luoyexuge/article/details/49104583>)

现在回想学pandas很意外，记得五月份的时候刚刚跑到现在的公司实习，那个时候公司没发电脑，当时天天去那就看书（自己的电脑被朋友拿走了），就这样看了半个月的pandas，后来也就慢慢用上了，下面是pa...



luoyexuge (<http://blog.csdn.net/luoyexuge>) 2015年10月13日 21:19 7829


## pandas做数据分析(三):常用预处理操作 (<http://blog.csdn.net/xierhacker/article/details/6593...>)

数据预处理常见的pandas实现

 xierhacker (<http://blog.csdn.net/xierhacker>) 2017年05月02日 20:18 1311

## 利用Python数据分析：数据规整化（五）(<http://blog.csdn.net/kkbb8811/article/details/605...>)


import pandas as pd from pandas import Series,DataFrame import numpy as np import re # 计算指标/哑变量 (dum...

 kkbb8811 (<http://blog.csdn.net/kkbb8811>) 2017年03月06日 13:38 208

4


## 基于【pandas】的数据预处理（含定制【OneHotEncoder】方法）(<http://blog.csdn.net/lixi...>)

20161130 - 实验实录，目前完成了训练集、测试集 7:3 分割，category特征重组及其哑编码，numeric特征噪声清洗及其标准化，可视化方法定制、基于pandas的OneHotEnco...

 lixiaowang\_327 ([http://blog.csdn.net/lixiaowang\\_327](http://blog.csdn.net/lixiaowang_327)) 2016年11月29日 14:45 3611


## One-Hot Encoding独热编码(<http://blog.csdn.net/wang4959520/article/details/50973137>)

机器学习 数据预处理之独热编码（One-Hot Encoding）问题由来 在很多机器学习任务中，特征并不总是连续值，而有可能是分类值。例如，考虑一下的三个特征： ...

 wang4959520 (<http://blog.csdn.net/wang4959520>) 2016年03月24日 17:01 1681


## 所涉及到的几种 sklearn 的二值化编码函数：OneHotEncoder(), LabelEncoder(), LabelBinar...

转自<http://blog.csdn.net/haramshen/article/details/53169963> 所涉及到的几种 sklearn 的二值化编码函数：OneHotEncoder(...


 gao1440156051 (<http://blog.csdn.net/gao1440156051>) 2017年02月14日 11:43 1788


## pandas中利用get\_dummies()进行独热编码（One-Hot encoding）([http://blog.csdn.net/wl\\_...](http://blog.csdn.net/wl_...))

在机器学习分类任务中，经常存在一个特征有多个分类变量值，例如在kaggle中的Titanic比赛数据中，Embarked的值有S,C,Q。我们这个时候要对离散型数据进行onehot编码处理，至于one...


 wl\_ss ([http://blog.csdn.net/wl\\_ss](http://blog.csdn.net/wl_ss)) 2017年11月11日 20:57 189


## Python中的虚拟变量(dummy variables) (<http://blog.csdn.net/lll1528238733/article/details/...>)

 虚拟变量(dummy variables) 虚拟变量，也叫哑变量和离散特征编码，可用来表示分类变量、非数量因素可能产生的影响。① 离散特征的取值之间有大小的意义 例如：尺寸(L、XL、XXL) ...

 lll1528238733 (<http://blog.csdn.net/lll1528238733>) 2017年07月14日 14:26 2812

## 2017.08.01回顾 xgboost get\_dummies 尽可能留变量 (<http://blog.csdn.net/strwolf/article/d...>)

 节奏不要乱，做应该做的事情，一些不应该做的事情需要克制！1、上午就是抓取建模变量，对变量做一些初步encode的工作，集中注意力弄了一上午，应该完成了1/3 2、继续zillow数据建模，有几...

 strwolf (<http://blog.csdn.net/strwolf>) 2017年08月02日 10:42 196

## pandas将类别属性转化为数值属性的方法 (<http://blog.csdn.net/chenpe32cp/article/details/7...>)

原文地址 离散特征的编码分为两种情况：1、离散特征的取值之间没有大小的意义，比如color : [red,blue],那么就使用one-hot编码 2、离散特征的取值有大小的意义，比如size:[...

 chenpe32cp (<http://blog.csdn.net/chenpe32cp>) 2017年07月19日 22:48 928

## 大神手把手教你：(Python)序列数据的One Hot编码 (<http://blog.csdn.net/Datapad/article/de...>)

原文链接机器学习算法无法直接用于数据分类。数据分类必须转换为数字才能进一步进行。在本教程中，你将发现如何将输入或输出的序列数据转换为一种热编码，以便于你在Python中深度学习的序列分类问题中使用。看...



Datapad (<http://blog.csdn.net/Datapad>) 2017年09月02日 15:50 409

## Python利用pandas/sklearn处理类别型特征；手动分段与离散化处理连续型特征 (<http://blog.csdn.net/Datapad/article/details/52836051>)

类别型变量特征：独热向量编码/One-Hot-Encoding (Dummy variables) 颜色：红、黄、紫[1,0,0] [0,1,0] [0,0,1] LR =...



cymy001 (<http://blog.csdn.net/cymy001>) 2017年11月19日 19:12 320

4

## Python序列类型常用函数练习：enumerate() reversed() sorted() zip() (<http://blog.csdn.net/u011583927/article/details/52836051>)

本文使用代码示例，练习使用序列类型的常用函数练习，包括：enumerate() reversed() sorted() zip()...



u011583927 (<http://blog.csdn.net/u011583927>) 2016年12月08日 15:58 686



## 机器学习笔记——Scikit-learn库中的数据预处理：独热编码（二） ([http://blog.csdn.net/m0\\_37324740/article/details/52836051](http://blog.csdn.net/m0_37324740/article/details/52836051))

在上一篇博客中介绍了数值型数据的预处理但是真实世界的数据集通常都含有分类型变量(categorical value)的特征。当我们讨论分类型数据时，我们不区分其取值是否有序。比如T恤尺寸是有序的，因...



m0\_37324740 ([http://blog.csdn.net/m0\\_37324740](http://blog.csdn.net/m0_37324740)) 2017年08月14日 20:51 436

## Pandas 文本数据方法 get\_dummies() (<http://blog.csdn.net/claroja/article/details/64923408>)

Series.str.get\_dummies(sep='|')分割字符串并返回各个字符的复杂矩阵 参数: sep : 字符串, 默认为'|' 返回值: 数据框(DataFrame)>>> S...



claroja (<http://blog.csdn.net/claroja>) 2017年03月22日 18:35 276