

null的专栏

Keep your eyes open and your feet moving forward. You'll find what you need.

个人声明

欢迎大家加群，探讨与机器学习相关技术相关的话题：
101620539
博客的主要内容主要是自己的学习笔记，并结合个人的理解，供各位在学习过程中参考，若有疑问，欢迎提出；若有侵权，请告知博主删除，原创文章转载还请注明出处。

我写的书：

购买链接：
京东-Python机器学习算法

个人资料



zhiyong_will

+ 加关注

发私信



≡

恒

访问：788977次
积分：8316
等级：

BLOG > 6

排名：第2239名

原创：143篇
转载：1篇
译文：1篇
评论：402条

深度学习Deep Learning



文章：8篇
阅读：27689

优化算法

文章：14篇

 **CSDN日报20170707——《稀缺：百分之二的选择》** [征文 | 你会为 AI 转型么？](#) [每周荐书 | Android、Keras、ES6（评论送书）](#)

原 数据处理——One-Hot Encoding

标签：[数据处理](#) [One-Hot](#)

2015-03-03 16:54 22196人阅读 评论



微信关注CSDN
获得无限技术资源

快速回复

我要收藏

分类：[论文与材料的学习笔记（32）](#)

版权声明：本文为博主原创文章，未经博主允许不得转载。

目录(?)	[+]
目录(?)	[+]

一、One-Hot Encoding

One-Hot编码，又称为一位有效编码，主要是采用 N 位状态寄存器来对 N 个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候只有一位有效。

在实际的机器学习的应用任务中，特征有时候并不总是连续值，有可能是一些分类值，如性别可分为“male”和“female”。在机器学习任务中，对于这样的特征，通常我们需要对其进行特征数字化，如下面的例子：
有如下三个特征属性：

- 性别：["male", "female"]
- 地区：["Europe", "US", "Asia"]
- 浏览器：["Firefox", "Chrome", "Safari", "Internet Explorer"]

对于某一个样本，如["male", "US", "Internet Explorer"]，我们需要将这个分类值的特征数字化，最直接的方法，我们可以采用序列化的方式：[0,1,3]。但是这样的特征处理并不能直接放入机器学习算法中。

二、One-Hot Encoding的处理方法

对于上述的问题，性别的属性是二维的，同理，地区是三维的，浏览器则是思维的，这样，我们可以采用One-Hot编码的方式对上述的样本["male", "US", "Internet Explorer"]”编码，“male”则对应着[1, 0]，同理“US”对应着[0, 1, 0]，“Internet Explorer”对应着[0,0,0,1]。则完整的特征数字化的结果为：[1,0,0,1,0,0,0,0,1]。这样导致的一个结果就是数据会变得非常的稀疏。

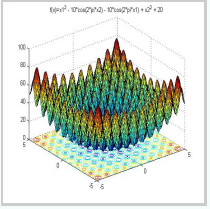
三、实际的Python代码

```
[python]

01. from sklearn import preprocessing
02.
03. enc = preprocessing.OneHotEncoder()
04. enc.fit([[0,0,3],[1,1,0],[0,2,1],[1,0,2]])
05.
06. array = enc.transform([[0,1,3]]).toarray()
07.
08. print array
```

```
[python]

01. from sklearn import preprocessing
02.
03. enc = preprocessing.OneHotEncoder()
04. enc.fit([[0,0,3],[1,1,0],[0,2,1],[1,0,2]])
05.
```



阅读：80226

机器学习，数据挖掘算法

文章：42篇

阅读：371911

文章分类

- Machine Learning (57)
- Deep Learning (10)
- Optimization Algorithm (15)
- 论文与材料的学习笔记 (33)
- Computational Advertising (3)
- NLP (3)
- Recommender System (4)
- 每周算法练习 (9)
- Data Structure & Algorithm (6)
- 设计模式 (1)
- Hadoop (4)
- Python (13)
- C/C++ (8)
- Matlab技巧 (2)
- Spark (1)
- Linux (3)
- PHP (2)
- Practice in Job (1)
- Web Spider (1)

联系我

Email：
zhaozhiyong1989@126.com

```
06.         array = enc.transform([[0,1,3]]).toarray()
07.
08.     print array
```

结果：[[1. 0. 0. 1. 0. 0. 0. 0. 1.]]



顶10

踩0

- 上一篇 优化算法——凸优化的概述
- 下一篇 CTR——人工神经网络+决策树

相关文章推荐

- 数据处理——One-Hot Encoding
 - 海量数据处理系列——BloomFilter
 - BloomFilter——大规模数据处理利器
 - postgresql集群方案hot standby初级测试(二) ——...
 - 机器学习系列：（三）特征提取与处理
- BloomFilter——大规模数据处理利器[转]
 - 码农周刊分类整理
 - 海量数据处理系列——C语言下实现bitmap算法(转)
 - 高负载高并发网站架构分析
 - 海量数据处理专题4——堆

猜你在找

- 机器学习之概率与统计推断
 - 机器学习之凸优化
 - 响应式布局全新探索
 - 深度学习基础与TensorFlow实践
 - 前端开发在线峰会
- 机器学习之数学基础
 - 机器学习之矩阵
 - 探究Linux的总线、设备、驱动模型
 - 深度学习之神经网络原理与实战技巧
 - TensorFlow实战进阶：手把手教你做图像识别应用

查看评论

2楼 [zhiyong_will](#) 2017-04-07 09:51发表



回复wanzi_antang：fit中的四个代表四个训练样本，表示的是每一维的可能取值，比如“性别”这一维只有两个值：0和1，那么在四个样本中第一维是0或者1，其他依次类推

Re: [wanzi_antang](#) 2017-04-07 20:08发表



回复google19890102：谢谢 看明白了

1楼 [大号小白兔](#) 2017-03-09 10:30发表



请问enc.fit([[0,0,3],[1,1,0],[0,2,1],[1,0,2]]) 里面的数组代表什么呢？

Re: [xbbbbb](#) 2017-04-21 15:08发表



回复a1b2c3d4123456：仔细看了下原文，其实就是训练数据，在这个例子里面就是随便组合的几组（学习到 [len(性别), len(地区), len(浏览器)], 而不是直接给出来）

Re: [xbbbbb](#) 2017-04-21 15:09发表



回复a1b2c3d4123456：训练数据集，可以计算出[len(性别),len(地区),len(浏览器)]

Re: [xbbbbb](#) 2017-04-21 15:06发表



回复a1b2c3d4123456：仔细看了下原文，其实就是训练数据，在这个例子里面就是随便组合的几组（学习到 [len(性别), len(地区), len(浏览器)], 而不是直接给出来）

浏览..

上传图片

Re: zhiyong_will 2017-03-09 10:55发表 



回复a1b2c3d4123456：这个只是几个数据集，表明的是每一维的范围，具体可以参考：<http://scikit-learn.org/stable/modules/preprocessing.html#preprocessing>

您还没有登录,请[\[登录\]](#)或[\[注册\]](#)

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

文章存档

- 2017年06月 (4)
- 2017年05月 (1)
- 2017年04月 (1)
- 2017年03月 (4)
- 2017年02月 (4)

 展开

阅读排行

- 简单易学的机器学习算法 (42995)
- 简单易学的机器学习算法 (31718)
- 简单易学的机器学习算法 (24051)
- 数据处理——One-Hot Encoding (22166)
- Python技巧——list与字典 (21417)
- python——时间与时间戳 (19983)
- 简单易学的机器学习算法 (19570)
- 简单易学的机器学习算法 (17076)
- 优化算法——人工蜂群算法 (15786)
- 简单易学的机器学习算法 (15274)

评论排行	
简单易学的机器学习算法	(55)
简单易学的机器学习算法	(27)
优化算法——人工蜂群算	(25)
推荐算法——基于矩阵分	(22)
简单易学的机器学习算法	(22)
简单易学的机器学习算法	(17)
简单易学的机器学习算法	(15)
社团划分——Fast Unfolc	(15)
机器学习算法实践——K-	(13)
论文中的机器学习算法—	(11)

最新评论	
《Python机器学习算法》的写作/ zhiyong_will: @qq_15618989:感谢支持	
《Python机器学习算法》的写作/ qq_15618989: 京东上已经有这本书了	
推荐算法——基于矩阵分解的推 zhiyong_will: @wzx19840423:那是正则化的部分	
推荐算法——基于矩阵分解的推 zhiyong_will: @wzx19840423:你和的是整个矩阵，只是取每一个位置上的值	
推荐算法——基于矩阵分解的推 强迫症专用头像: 请问一下，梯度的值为什么还要乘以qk,j和Pi,k这些？	
推荐算法——基于矩阵分解的推 强迫症专用头像: 楼主，请问一下： $p = p + \alpha * (2 * error * q - \beta * p)$...	
推荐算法——基于矩阵分解的推 强迫症专用头像: 楼主，请问下值只能一个一个的拟合吗？能不能一行一行的拟合，或者直接拟合整个矩阵？	
简单易学的机器学习算法——las KaaiLee: 好像没给数据集呢	
简单易学的机器学习算法——因 ww_up: 请问训练和测试数据对应的是机器学习实战上的什么数据？	
机器学习中的常见问题——几种 X_qiu: 不错 学习了 谢谢	

其他人的博客	
Rachel Zhang的专栏	
结构之法 算法之道	
美团点评技术团队	

公司简介 | 招贤纳士 | 广告服务 | 联系方式 | 版权声明 | 法律顾问 | 问题报告 | 合作伙伴 | 论坛反馈

 网站客服  杂志客服  微博客服  webmaster@csdn.net  400-660-0108 | 北京创新乐知信息技术有限公司 版权所有 | 江苏知之为计算机有限公司 |

江苏乐知网络技术有限公司

京 ICP 证 09002463 号 | Copyright © 1999-2017, CSDN.NET, All Rights Reserved 

公司简介 | 招贤纳士 | 广告服务 | 联系方式 | 版权声明 | 法律顾问 | 问题报告 | 合作伙伴 | 论坛反馈

 网站客服  杂志客服  微博客服  webmaster@csdn.net  400-660-0108 | 北京创新乐知信息技术有限公司 版权所有 | 江苏知之为计算机有限公司 |

江苏乐知网络技术有限公司

京 ICP 证 09002463 号 | Copyright © 1999-2017, CSDN.NET, All Rights Reserved 

