

CSDN新首页上线啦，邀请你来立即体验！(http://blog.csdn.net/)

立即体验

广告

立即体验

博客 (http://blog.csdn.net/?ref=toolbar)

学院 (http://edu.csdn.net/?ref=toolbar)

更多

下载 (http://www.csdn.net/download)

更多

搜索

搜索

登录 (https://passport.csdn.net/account/login?ref=toolbar)

注册 (http://passport.csdn.net/account/mobileregister?ref=toolbar&action=mobileRegister)

历史最详细的XGBoost实战（上）

原创

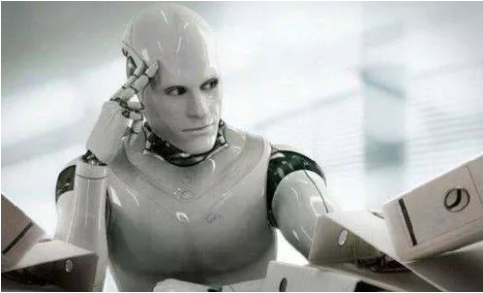
2017年10月31日 00:00:00

104

历史最详细的XGBoost实战（上）

作者：章华燕

编辑：祝鑫泉




零环境介绍:

- Python版本: 3.6.2
- 操作系统: Windows
- 集成开发环境: PyCharm

安装Python环境:

1. 安装Python:

首先，我们需要安装Python环境。本人选择的是64位版本的Python 3.6.2。去Python官网<https://www.python.org> /选择相应的版本并下载。如下如所示:



python-3.6.2-amd64.exe

接下来安装，并最终选择将Python加入环境变量中。

2. 安装依赖包:

可以去网址:<http://www.lfd.uci.edu/~gohlke/pythonlibs/>中去下载你所需要的如下Python安装包:

numpy-1.13.1+mk1-cp36-cp36m-win_amd64.whl

scipy-0.19.1-cp36-cp36m-win_amd64.whl xgboost-0.6-cp36-cp36m-win_amd64.whl

假设上述三个包所在的目录为D:\Application，则运行Windows 命令行运行程序cmd，并将当前目录转到这两个文件所在的目录下。并依次执行如下操作安装这两个包:

```
>> pip install numpy-1.13.1+mk1-cp36-cp36m-win_amd64.whl
>> pip install scipy-0.19.1-cp36-cp36m-win_amd64.whl
>> pip install xgboost-0.6-cp36-cp36m-win_amd64.whl
```

3. 安装Scikit-learn

众所周知，scikit-learn是Python机器学习最著名的开源库之一。因此，我们需要安装此库。执行如下命令安装scikit-learn机器学习库:

```
>> pip install -U scikit-learn
```

4. 测试是否安装成功

```
from sklearn import svm
X = [[0, 0], [1, 1]]>>> y = [0, 1]
clf = svm.SVC() >>> clf.fit(X, y)
clf.predict([[2., 2.]]) array([1])
import xgboost as xgb
```

他的最新文章

更多文章 (http://blog.csdn.net/szm21c11u68n04vdcLmJ)

原创

粉丝

喜欢

未开通

42

1

1

(https://gitee.com/szm21c11u68n04vdcLmJ)

他的最新文章

更多文章 (http://blog.csdn.net/szm21c11u68n04vdcLmJ)

机器学习从零开始系列连载(2)——线性回归 (http://blog.csdn.net/SzM21c11u68n04vdcLmJ/article/details/78651353)

无人车之父Sebastian Thrun: 技术小白，也能从零开始造一辆无人车！ (http://blog.csdn.net/SzM21c11u68n04vdcLmJ/article/details/78651350)

简单易懂的自动编码器 (http://blog.csdn.net/SzM21c11u68n04vdcLmJ/article/details/78635586)

Qualcomm

Unable to Conn

The Proxy was unable to connect to the remote site. responding to requests. If you feel you have reached please submit a ticket via the link provided below.

URL: http://pos.baidu.com/s?hei=250&wid=300&di=u%2Fblog.csdn.net%2FSzM21c11u68n04vdcLmJ%2F

在线课程

腾讯云容器服务架构实现介绍 (0)

讲师: 董晓杰

机器学习在5G网络中的应用 (http://edu.csdn.net/course/73?utm_source=blog9)

机器学习在5G网络中的应用 (http://edu.csdn.net/course/73?utm_source=blog9)

他的热门文章

人工智能到底有多火，年薪 25 万只是白菜价..... (http://blog.csdn.net/szm21c11u68n04vdcLmJ/article/details/78410206)

223

Logistic回归实战篇之预测病马死亡率（三） (http://blog.csdn.net/szm21c11u68n04vdcLmJ/article/details/78307965)

187

机器学习损失函数、L1-L2正则化的前世今生 (http://blog.csdn.net/szm21c11u68n04vdcLmJ/article/details/78138887)

157

机器学习之——自动求导 (http://blog.csdn.net/szm21c11u68n04vdcLmJ/article/details/78188600)

157

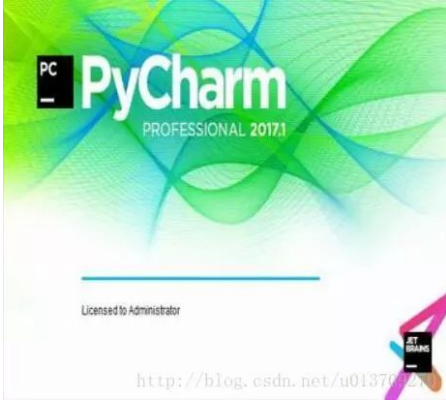
[视频讲解]史上最全面的正则化技术总结与分析！ (http://blog.csdn.net/szm21c11u68n04vdcLmJ/article/details/78188600)

第1页 共7页

2017/11/30 下午4:46

击 下一步 就行了。

广告04vdcLmJ/article/details/78547038)
157



注意：PyCharm软件是基于Java开发的，所以安装该集成开发环境前请先安装JDK，建议安装JDK1.8。

经过上述步骤，基本上软件环境的问题全部解决了，接下来就是实际的XGBoost库实战了.....

二

XGBoost的优点

1. 正则化

XGBoost在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数、每个叶子节点上输出的score的L2模的平方和。从Bias-varianctradeoff角度来讲，正则项降低了模型的variance，使学习出来的模型更加简单，防止过拟合，这也是xgboost优于传统GBDT的一个特性。

2. 并行处理

XGBoost工具支持并行。Boosting不是一种串行的结构吗？怎么并行的？注意XGBoost的并行不是tree粒度的并行，XGBoost也是一次迭代完才能进行下一次迭代的（第t次迭代的代价函数里包含了前面t-1次迭代的预测值）。XGBoost的并行是在特征粒度上的。

我们知道，决策树的学习最耗时的一个步骤就是对特征的值进行排序（因为要确定最佳分割点），XGBoost在训练之前，预先对数据进行了排序，然后保存为block结构，后面的迭代中重复地使用这个结构，大大减小计算量。这个block结构也使得并行成为了可能，在进行节点的分裂时，需要计算每个特征的增益，最终选增益最大的那个特征去做分裂，那么各个特征的增益计算就可以开多线程进行。

3. 灵活性

XGBoost支持用户自定义目标函数和评估函数，只要目标函数二阶可导就行。

4. 缺失值处理

对于特征的值有缺失的样本，xgboost可以自动学习出它的分裂方向。

5. 剪枝

XGBoost 先从上到下建立所有可以建立的子树，再从底到顶反向进行剪枝。比起GBM，这样不容易陷入局部最优解。

6. 内置交叉验证

XGBoost允许在每一轮boosting迭代中使用交叉验证。因此，可以方便地获得最优boosting迭代次数。而GBM使用网格搜索，只能检测有限个值。

三

XGBoost详解：

1. 数据格式

XGBoost可以加载多种数据格式的训练数据：

- 1. **libsvm** 格式的文本数据；
- 2. **Numpy** 的二维数组；
- 3. **XGBoost** 的二进制的缓存文件。加载的数据存储在对象 **DMatrix** 中。

下面一一列举：

- 加载libsvm格式的数据
>>> dtrain1 = xgb.DMatrix('train.svm.txt')
- 加载二进制的缓存文件
>>> dtrain2 = xgb.DMatrix('train.svm.buffer')

内容举报
返回顶部

广告

- 加载numpy的数组

```
>>> data = np.random.rand(5,10) # 5 entities, each contains 10 features
>>> label = np.random.randint(2, size=5) # binary target
>>> dtrain = xgb.DMatrix( data, label=label)
```
- 将scipy.sparse格式的数据转化为 **DMatrix** 格式

```
>>> csr = scipy.sparse.csr_matrix( (data, (row,col)) ) >>> dtrain = xgb.DMatrix( csr )
```
- 将 DMatrix 格式的数据保存成XGBoost的二进制格式，在下次加载时可以提高加载速度，使用方式如下

```
>>> dtrain = xgb.DMatrix('train.svm.txt')
>>> dtrain.save_binary("train.buffer")
```
- 可以用如下方式处理 DMatrix中的缺失值：

```
>>> dtrain = xgb.DMatrix( data, label=label, missing = -999.0)
```
- 当需要给样本设置权重时，可以用如下方式

```
>>> w = np.random.rand(5,1)
>>> dtrain = xgb.DMatrix( data, label=label, missing = -999.0, weight=w)
```

2. 参数设置

XGBoost使用key-value字典的方式存储参数：

```
params = {
    'booster': 'gbtree', 'objective': 'multi:softmax', # 多分类的问题 'num_class': 10,          # 类别数，与
multisoftmax 并用
    'gamma': 0.1,          # 用于控制是否后剪枝的参数,越大越保守，一般0.1、0.2这样子。
    'max_depth': 12,          # 构建树的深度，越大越容易过拟合 'lambda': 2,          # 控制模型复杂度的权重值
的L2正则化项参数，参数越大，模型越不容易过拟合。
    'subsample': 0.7,          # 随机采样训练样本
    'colsample_bytree': 0.7,          # 生成树时进行的列采样 'min_child_weight': 3, 'silent': 1,          # 设置
成1则没有运行信息输出，最好是设置为0.
    'eta': 0.007,          # 如同学习率
    'seed': 1000, 'nthread': 4,          # cpu 线程数}
```

3. 训练模型

有了参数列表和数据就可以训练模型了

```
num_round = 10
bst = xgb.train( plst, dtrain, num_round, evallist )
```

4. 模型预测

```
# X_test类型可以是二维List，也可以是numpy的数组
dtest = DMatrix(X_test) ans = model.predict(dtest)
```

5. 保存模型

- 在训练完成之后可以将模型保存下来，也可以查看模型内部的结构

```
bst.save_model('test.model')
```
- 导出模型和特征映射（Map）
 你可以导出模型到txt文件并浏览模型的含义：

```
# dump model
bst.dump_model('dump.raw.txt')
# dump model with feature map
bst.dump_model('dump.raw.txt','featmap.txt')
```

6. 加载模型

通过如下方式可以加载模型：

```
bst = xgb.Booster({'nthread':4}) # init model
bst.load_model("model.bin")      # load data
```



内容举报

返回顶部

广告

好长长长累了吧？ 快给作者和编辑鼓掌！！
还有么？有！！
未完待续。。。
休息一下，坐等更新哇。

PS：
1. 考虑到文章太长，就给大家写成分段“函数”了。
2. 后续为大家讲解参数详解以及实战

关于本文adaboost 相关知识和其他问题
欢迎大家加群在群中探讨
欢迎留言或赞赏。

推
荐
阅
读

1. 客官，来嘛，谷歌小菜请你尝尝！ (http://mp.weixin.qq.com/s?__biz=MzUyMjE2MTE0Mw==&mid=2247484028&idx=1&sn=06a7d4c740470cf752ccf3010f37f146&chksm=f9d15ce4cea6d5f29eab6e13abafa7f5acb0067540202297eea506f26b234e6ab7d383c88299&scene=21#wechat_redirect)
2. 趣谈深度学习核心----激活函数 (http://mp.weixin.qq.com/s?__biz=MzUyMjE2MTE0Mw==&mid=2247484062&idx=1&sn=97019bf9aac37a653725008a1e697305f&chksm=f9d15c06cea6d51056f42ea79a37fe9fb65de596f2cbd19fdafc7a828deecb8b32c712e1248e&scene=21#wechat_redirect)
3. 朴素贝叶斯实战篇之新浪新闻分类 (http://mp.weixin.qq.com/s?__biz=MzUyMjE2MTE0Mw==&mid=2247484255&idx=1&sn=d101099b80d560c3f6bca1833f39c5fe&chksm=f9d15dc7cea6d4d1f39147ae5c63e5e8e456d9f393fec5441c6ba02b2b9c6358126578da9ddb&scene=21#wechat_redirect)
4. Object Detection R-CNN (http://mp.weixin.qq.com/s?__biz=MzUyMjE2MTE0Mw==&mid=2247484206&idx=1&sn=f9084165a4673affd8e23ac97f707eb8&chksm=f9d15db6cea6d4a04a777d45ae3e3f1a3ef6f657352c98db9664084fe4b4f802273dde7ac943&scene=21#wechat_redirect)

扫描个人微信号，
拉你进机器学习大牛群。
福利满满，名额已不多...



80%的AI从业者已关注我们微信公众号



版权声明：本文为博主原创文章，未经博主允许不得转载。

内容举报
返回顶部





相关文章推荐

广告

史上最详细的XGBoost实战（下） (http://blog.csdn.net/SzM21C11U68n04vdcLmJ/articl...

作者：章华燕编辑：田旭四 XGBoost 参数详解在运行XGboost之前，必须设置三种类型成熟：general parameters, booster parameters...

SzM21C11U68n04vdcLmJ (http://blog.csdn.net/SzM21C11U68n04vdcLmJ) 2017年11月12日 00:00 36

史上最详细的CocoaPods安装教程 (http://blog.csdn.net/qq_33236947/article/details/501...

什么是CocoaPods CocoaPods是OS X和iOS下的一个第三类库管理工具，通过CocoaPods工具我们可以为项目添加被称为“Pods”的依赖库（这些类库必须是CocoaPod...

qq_33236947 (http://blog.csdn.net/qq_33236947) 2015年12月03日 13:46 66



【免费技术直播】数据科学家，从入门到精进

数据科学家究竟是一群怎样的人？来自北美数据科学职场前线，为你带来作为数据科学家的第一手经验..

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjfdRHT0IZ0qnfK9ujYzP1nsrjD10Aw-5Hc3rHnYnHb0TAq15HfLPWRznb0T1Yym1f3PWcdm19-P1m4mHn30AwY5HDdnHcsn1DdPjT0lgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEpZNGXy-jULNzTvREuANYmy-_Q1mknHqdlAdxTvqdThP-5yF_UvTkn0KzujY4rHb0mhYqn0KsTWYs0ZNGUjYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1YknWbLP0)

史上最详细的webpack 讲解2（DefinePlugin中的淫技巧） (http://blog.csdn.net/sinat_1...

今天我突然发现我的掘金原创排行已经落到了120位，这是什么原因，因为我分享的不够多，还是我分享的不够好，看的人不多，又好几天没和大家见面了，来吧！死也死在分享上面，怎么说呢，今天讲解的东西也不是很深奥...

sinat_17775997 (http://blog.csdn.net/sinat_17775997) 2017年04月12日 10:46 7300

史上最详细的vsftpd配置文件讲解 (http://blog.csdn.net/weiyuefei/article/details/51564367)

本文根据RedKing的帖子整理节选而来。原文地址在http://bbs.51cto.com/thread-717151-1.html。vsftpd作为一个主打安全的FTP服务器，有很多的选项...

weiyuefei (http://blog.csdn.net/weiyuefei) 2016年06月02日 10:23 9012

史上最详细的Android原生APP中添加ReactNative 进行混合开发教程 (http://blog.csdn....

原文地址：http://www.jianshu.com/p/22aa14664cf9?open_source=weibo_search 转载过来，以备日后查看背景 React Native...

dodod2012 (http://blog.csdn.net/dodod2012) 2017年06月21日 16:51 346



程序员跨越式成长指南

完成第一次跨越，你会成为具有一技之长的开发者，月薪可能翻上几番；完成第二次跨越，你将成为拥有局部优势或行业优势的专业人士，获得个人内在价值的有效提升和外在收入的大幅跃迁...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjfdRjD0IZ0qnfK9ujYzP1f4PjnY0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YkPhDLujwhPH9bPAm4PWmk0AwY5HDdnHcsn1DdPHD0lgF_5y9YIZ0IQzqMpgwBUvqoQhP8QvGIAPCmgfEmvq_lyd8Q1R4uWc4uHf3uAckPHRkPWN9Ph5HDknWFBmhkEusKzujY4rHb0mhYqn0KsTWYs0ZNGUjYkPHTYn1mk0AqGujYkn10snj10APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1Y1PHnznf)

史上最详细的javascript闭包（Closure）说明 (http://blog.csdn.net/Darin_Zanyar/article...

闭包（closure）是JavaScript语言的一个难点，也是它的特色，很多高级应用都要依靠闭包实现。很早就接触过闭包这个概念了，但是一直糊里糊涂的，没有能够弄明白JavaScript的闭包到底是什...

Darin_Zanyar (http://blog.csdn.net/Darin_Zanyar) 2016年08月19日 16:00 82

史上最详细的Android Studio系列教程二--基本设置与运行 (http://blog.csdn.net/aa2061...

上面一篇博客，介绍了Studio的优点与1.0 RC的安装与上手体验，没想到google的更新速度这么快，已经出了RC 2版本，主要是修复一些bug。那么今天就带大家预览下Stduio的界面与基本功能...

aa20616012 (http://blog.csdn.net/aa20616012) 2015年09月06日 10:21 258

内容举报
返回顶部



广告

史上最详细的八个皇后算法解析【php版本】
(http://blog.csdn.net/zhengxiaojunkite/art...)
题目：八皇后问题是一个以国际象棋为背景的问题：如何能够在8×8的国际象棋棋盘上放置八个皇后，使得任何一个皇后都无法直接吃掉其他的皇后。为了达到此目的，任两个皇后都不能处于同一条横行、纵行或斜线上。 ...
zhengxiaojunkite (http://blog.csdn.net/zhengxiaojunkite) 2015年05月16日 16:23 342

史上最详细的LXR安装介绍
(http://blog.csdn.net/mosenyang/article/details/50755539)
史上最详细的LXR安装介绍（Ubuntu14.04+Apache2.4.7）简介：LXR（Linux Cross Reference）是一个通过交叉索引方便用户查看项目源代码的工具。项目地址：htt...
mosenyang (http://blog.csdn.net/mosenyang) 2016年02月27日 15:39 1478

史上最详细的Hashtable详解--源码分析
(http://blog.csdn.net/yan_wenliang/article/detail...)
史上最详细的Hashtable详解--源码分析
yan_wenliang (http://blog.csdn.net/yan_wenliang) 2016年03月29日 10:36 593


史上最详细的JavaScript事件使用指南
(http://blog.csdn.net/maodoudou1217/article/det...)
史上最详细的JavaScript事件使用指南事件流事件流描述的是从页面中接收事件的顺序，IE和Netscape提出来差不多完全相反的事件流的概念，IE事件流是事件冒泡流，N...
maodoudou1217 (http://blog.csdn.net/maodoudou1217) 2015年08月10日 13:59 253

史上最详细的struts 2 标签整理
(http://blog.csdn.net/cnboynet/article/details/6922709)
a a标签创建一个HTML超链接，等价于HTML的示范代码：登陆更多 a 信息 action 使用action标签 可以允许在JSP页面中直接调用Act...
cnboynet (http://blog.csdn.net/cnboynet) 2011年10月31日 20:39 380

史上最详细的Android Studio系列教程二--基本设置与运行
(http://blog.csdn.net/dykun_...)
原文链接：http://stormzhang.com/devtools/2014/11/28/android-studio-tutorial2/ 上面一篇博客，介绍了Studio的优点与1...
dykun_1225 (http://blog.csdn.net/dykun_1225) 2015年08月11日 15:34 182

史上最详细的WIN7下WIFI共享上网教程
(http://blog.csdn.net/binnygoal/article/details/7...)
1.打开WIN7开始菜单，在左下角的框中输入CMD，搜索出来的CMD.EXE对着它右键，选择以“管理员身份运行”。2。（1）netsh wlan set hostednetwork mode=...
binnygoal (http://blog.csdn.net/binnygoal) 2011年12月15日 20:41 389

史上最详细的CocoaPods安装教程
(http://blog.csdn.net/xiao19911130/article/details/50...)
虽然网上关于CocoaPods安装教程多不胜数,但是我在安装的过程中还是出现了很多错误,所以大家可以照下来步骤装一下,我相信会很好用. 前言在iOS项目中使用第三方类库可...
xiao19911130 (http://blog.csdn.net/xiao19911130) 2015年11月26日 17:23 175



服务器上Tomcat运行jsp项目与服务器上tomcat域名解析外网访问-...

http://download... 2017年07月28日 14:34 486KB 下载

史上最详细的iOS之事件的传递和响应机制-原理篇
(http://blog.csdn.net/u011363981/art...)
前言：按照时间顺序，事件的生命周期是这样的：事件的产生和传递（事件如何从父控件传递到子控件并寻找到最合适的view、寻找最合适的view的底层实现、拦截事件的处理）->找到最合适的vi...
u011363981 (http://blog.csdn.net/u011363981) 2017年05月12日 15:00 139

移植QT5.6到嵌入式开发板（史上最详细的QT移植教程）
(http://blog.csdn.net/lizuobin...)
目前网上的大多数QT移植教程还都停留在qt4.8版本，或者还有更老的Qtopia,但是目前Qt已经发展到最新的5.7版本了，我个人也已经使用了很长一段时间了的qt5.6 for w...
lizuobin2 (http://blog.csdn.net/lizuobin2) 2016年09月28日 08:54 20627

【Caffe安装】caffe安装系列——史上最详细的安装步骤 (http://blog.csdn.net/haoji007/...



广告

说明网上关于caffe的安装教程非常多，但是关于每一步是否操作成功，出现了什么样的错误又该如何处理没有给出说明。因为大家的操作系统的环境千差万别，按照博客中的教程一步步的安装，最后可能失败——这是...

 haoji007 (http://blog.csdn.net/haoji007) 2016年07月31日 22:14  42499

史上最详细的解决 Amoeba连接mysql出错 解决方案 (http://blog.csdn.net/XMZ_JAVA/ar...

今天配置mysql的主从复制 用到了Amoeba。从安装到启动服务，我深深的感受到这个世界的恶意。首先是安装错误的解决，连接错误的兄弟可以直接往下拉。 1.出现 JAVA_HOME enviro...

 XMZ_JAVA (http://blog.csdn.net/XMZ_JAVA) 2017年02月07日 17:06  1338

