

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



博客 (//blog.csdn.net/?ref=toolbar) 学院 (//edu.csdn.net/?ref=toolbar)
下载 (//download.csdn.net/?ref=toolbar) GitChat (//gitbook.cn/?ref=csdn)
更多 ▾



weixin_3...
(//write.blog.csdn.net/postedita...
ref=toolbar)source=csdnblog

XGBoost : 二分类问题

翻译 2015年07月01日 15:38:23 标签 : xgboost (http://so.csdn.net/so/search/s.do?q=xgboost&t=blog)

16328

二分类问题

本文介绍XGBoost的命令行使用方法。Python和R的使用方法见
https://github.com/dmlc/xgboost/blob/master/doc/README.md (../doc/README.md)。
下面将介绍如何利用XGBoost解决二分类问题。以下使用的数据集见mushroom dataset
(https://archive.ics.uci.edu/ml/datasets/Mushroom)

简介

产生输入数据

XGBoost的输入数据格式和LibSVM一样。下面是XGBoost使用的输入数据格式：

| | | | |
|---|-----|---------|---------------------|
| 1 | 1 | 101:1.2 | 102:0.03 |
| 2 | 0 | 1:2.1 | 10001:300 10002:400 |
| 3 | ... | | |

每行表示一个样本，第一列的数字表示类别标签，表示样本所属于的类别，‘101’和‘102’表示特征索引，‘1.2’和‘0.03’是特征所对应的值。在二分类中‘1’表示正类，‘0’表示负类。同时类别标签支持概率标签，取值范围为[0,1]，表示样本属于某个类别的可能性。



工业设计公司

+ 关注

(http://blog.csdn.net/zc02051126)

码云

未开通

原创 132 粉丝 114 喜欢 0
(https://gitee.com/zc02051126)utm_source=csdn

他的最新文章

更多文章 (http://blog.csdn.net/zc02051126)

More Effective C++在leveldb中的体现
(http://blog.csdn.net/zc02051126/article/details/77844108)

梯度下降法概述 (http://blog.csdn.net/zc02051126/article/details/72845606)

笔记本 (http://blog.csdn.net/zc02051126/article/details/71919328)

立即体验



内容举报



返回顶部



第一步需要将数据集转化成libSVM形式，执行如下脚本


```
1 python mapfeat.py
2 python mknfold.py agaricus.txt 1
```

mapfeat.py和mknfold.py分别如下

```
1 #!/usr/bin/python
2 def loadfmap( fname ):
3     fmap = {}
4     nmap = {}
5     for l in open( fname ):
6         arr = l.split()
7         if arr[0].find('.') != -1:
```




在线课程



0

腾讯云容器服务架构介绍 ()

讲师：董晓杰



容器技术在58同城的实践 (http://edu.csdn.net/huiyiutm_source=blog9utm_source=blog9)

他的热门文章

- XGBoost : 在Python中使用XGBoost (http://blog.csdn.net/zc02051126/article/details/46771793)

66262
- XGBoost : 参数解释 (http://blog.csdn.net/zc02051126/article/details/46711047)


49276
- XGBoost : 二分类问题 (http://blog.csdn.net/zc02051126/article/details/46709599)

16301


绝对牛逼的t-SNE介绍 (http://blog.csdn.net/zc02051126/article/details/46709599)



XGBoost : 二分类问题 (http://blog.csdn.net/zc02051126/article/details/46709599)



内容举报



返回顶部



```

8     idx = int( arr[0].strip('.') )
9     assert idx not in fmap
10    fmap[ idx ] = {}
11    ftype = arr[1].strip(':')
12    content = arr[2]
13    else:
14        content = arr[0]
15    for it in content.split(','):
16        if it.strip() == '':
17            continue
18        k , v = it.split('=')
19        fmap[ idx ][ v ] = len(nmap)
20        nmap[ len(nmap) ] = ftype+'='+k
21    return fmap, nmap
22
23    def write_nmap( fo, nmap ):
24        for i in range( len(nmap) ):
25            fo.write('%d\t%s\t\n' % (i, nmap[i]))
26    # start here
27    fmap, nmap = loadfmap( 'agaricus-lepiota.fmap' )
28    fo = open( 'featmap.txt', 'w' )
29    write_nmap( fo, nmap )
30    fo.close()
31    fo = open( 'agaricus.txt', 'w' )
32    for l in open( 'agaricus-lepiota.data' ):
33        arr = l.split(',')
34        if arr[0] == 'p':
35            fo.write('1')
36        else:
37            assert arr[0] == 'e'
38            fo.write('0')
39        for i in range( 1,len(arr) ):
40            fo.write( ' %d:1' % fmap[i][arr[i].strip()] )
41        fo.write('\n')
42    fo.close()

```

```

1    #!/usr/bin/python
2    import sys
3    import random
4    if len(sys.argv) < 2:
5        print ('Usage:<filename> <k> [nfold = 5]')
6        exit(0)
7    random.seed( 10 )

```



XGBoost : 参数解释 (<http://blog.csdn.net/zc02051126/article/details/46711047>)

xgboost 二分类问题实例 (<http://blog.csdn.net/shenxiaoming77/article/details/76037930>)

xgboost使用案例二 (<http://blog.csdn.net/hb707934728/article/details/70739382>)

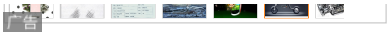
内容举报

返回顶部





```
1  / random.seed( 10 )
2
3  k = int( sys.argv[2] )
4
5  if len(sys.argv) > 3:
6      nfold = int( sys.argv[3] )
7  else:
8      nfold = 5
9
10 fi = open( sys.argv[1], 'r' )
11 ftr = open( sys.argv[1]+'train', 'w' )
12 fte = open( sys.argv[1]+'test', 'w' )
13
14 for l in fi:
15     if random.randint( 1 , nfold ) == k:
16         fte.write( l )
17     else:
18         ftr.write( l )
19
20 fi.close()
21 ftr.close()
22 fte.close()
```



运行完以上两个Python脚本将会产生训练数据集：'agaricus.txt.train' 和测试数据集：'agaricus.txt.test'

训练

执行如下命令行完成模型训练:

| | |
|---|-----------------------|
| 1 | xgboost mushroom.conf |
|---|-----------------------|

mushroom.conf文件用于配置训练模型和测试模型时需要的信息。每行的配置信息格式为：[attribute]=[value]：

```
1  # General Parameters, see comment for each definition
2
3  # can be gbtrees or gblines
4  booster = gbtrees
5
6  # choose logistic regression loss function for binary classification
7  objective = binary:logistic
8
9  # Tree Booster Parameters
10 # step size shrinkage
11 eta = 1.0
12
13 # minimum loss reduction required to make a further partition
14 gamma = 1.0
15
16 # minimum sum of instance weight(hessian) needed in a child
17 min_child_weight = 1
18
19 # maximum depth of a tree
20 max_depth = 3
```



2



内容举报



返回顶部



```
15 max_depth = 3
16
17 # Task Parameters
18 # the number of round to do boosting
19 num_round = 2
20 # 0 means do not save any model except the final round model
21 save_period = 0
22 # The path of training data
23 data = "agaricus.txt.train"
24 # The path of validation data, used to monitor training process, here [test] sets name of the validation set
25 eval[test] = "agaricus.txt.test"
26 # The path of test data
27 test:data = "agaricus.txt.test"
```

这里的booster采用gbtree，目标函数采用logistic regression。这意味着可以采用经典的梯度提升回归树进行计算（GBRT）。这种方法能够很好的处理二分类问题

以上的配置文件中给出了最常用的配置参数。如果想了解更多的参数，详见<https://github.com/dmlc/xgboost/blob/master/doc/parameter.md> (<https://github.com/dmlc/xgboost/blob/master/doc/parameter.md>)。如果不想在配置文件中配置算法参数，可以通过命令行配置，如下

```
1 xgboost mushroom.conf max_depth=6
```

这表示max_depth参数将被设置为6而不是配置文件中的3。当使用命令行参数时确保max_depth=6为一个参数，即参数之间不要含有间隔。如果既使用配置又使用命令行参数，则命令行参数会覆盖配置文件参数，即优先使用命令行参数

在以上的例子中使用tree booster计算梯度提升。如果想使用linear booster进行回归计算，可以修改booster参数为gblinear，配置文件中的其它参数都不需要修改，配置文件信息如下



```
1 # General Parameters
2 # choose the linear booster
3 booster = gblinear
4 ...
5
6 # Change Tree Booster Parameters into Linear Booster Parameters
7 # L2 regularization term on weights, default 0
8 lambda = 0.01
9 # L1 regularization term on weights, default 0
10 f ``agaricus.txt.test.buffer`` exists, and automatically loads from binary buffer if possible, this can speedup training p
```



```
11 - Buffer file can also be used as standalone input, i.e if buffer file exists, but original agaricus.txt.test was removed, xgbo
12 * Deviation from LibSVM input format: xgboost is compatible with LibSVM format, with the following minor differences:
13 - xgboost allows feature index starts from 0
14 - for binary classification, the label is 1 for positive, 0 for negative, instead of +1,-1
15 - the feature indices in each line *do not* need to be sorted
16 alpha = 0.01
17 # L2 regularization term on bias, default 0
18 lambda_bias = 0.01
19
20 # Regression Parameters
21 ...
```

预测

在训练好模型之后，可以对测试数据进行预测，执行如下脚本

```
1 xgboost mushroom.conf task=pred model_in=0003.model
```

对于二分类问题预测的输出结果为[0,1]之间的概率值，表示样本属于正类的概率。

模型展示

目前这还是个基本功能，只支持树模型的展示。XGBoost可以用文本的显示展示树模型，执行以下脚本

```
1 ../../xgboost mushroom.conf task=dump model_in=0003.model name_dump=dump.raw.txt
2 ../../xgboost mushroom.conf task=dump model_in=0003.model fmap=featmap.txt name_dump=dump.nice.txt
```

0003.model将会输出到dump.raw.txt和dump.nice.txt中。dump.nice.txt中的结果更容易理解，因为其中使用了特征映射文件featmap.txt

featmap.txt的格式为 featmap.txt: <featureid> <featurename> <q or i or int>\n :

- Feature id从0开始直到特征的个数为止，从小到大排列。
- i表示是二分类特征
- q表示数值变量，如年龄，时间等。q可以缺省
- int表示特征为整数(when int is hinted, the decision boundary will be integer)

计算过程监测

当运行程序时，会输出如下运行信息

⚠
内容举报

⬆
返回顶部



工业设计公司



广告



2



```
1 tree train end, 1 roots, 12 extra nodes, 0 pruned nodes ,max_depth=3
2 [0] test-error:0.016139
3 boosting round 1, 0 sec elapsed
4
5 tree train end, 1 roots, 10 extra nodes, 0 pruned nodes ,max_depth=3
6 [1] test-error:0.000000
```

计算过程中模型评价信息输出到错误输出流stderr中，如果希望记录计算过程中的模型评价信息，可以执行如下脚本

```
1 xgboost mushroom.conf 2>log.txt
```

在log.txt文件中记录如下信息

```
1 [0] test-error:0.016139
2 [1] test-error:0.000000
```

也可以同时监测训练过程和测试过程中的统计信息，可以通过如下方式进行配置

```
1 eval[test] = "agaricus.txt.test"
2 eval[trainname] = "agaricus.txt.train"
```

运行以上的脚本后得到的信息如下

```
1 [0] test-error:0.016139 trainname-error:0.014433
2 [1] test-error:0.000000 trainname-error:0.001228
```

运行规则是[name-printed-in-log] = filename，filename文件将会被加入检测进程并在每个迭代过程中对模型进行评价。

XGBoost同时支持多种统计量的监测，假设希望监测在训练过程每次预测的平均log-likelihood，只需要在配置文件中添加配置信息 eval_metric=logloss。再次运行log文件中将会有如下信息

```
1 [0] test-error:0.016139 test-negllik:0.029795 trainname-error:0.014433 trainname-negllik:0.027023
2 [1] test-error:0.000000 test-negllik:0.000000 trainname-error:0.001228 trainname-negllik:0.002457
```

内容举报

返回顶部

2











保存运行过程中的模型

如果现在运行过程中每两步保存一个模型，则可以设置参数`set save_period=2`。在当前文件夹将会看到模型0002.model。如果想修改模型输出的路径，则可以通过参数`dir=foldername`修改。缺省情况下XGBoost将会保持上次迭代的结果模型。

从已有模型继续计算

如果想从已有的模型继续训练，例如从0002.model继续计算，则用如下命令行

```
1 xgboost mushroom.conf model_in=0002.model num_round=2 model_out=continue.model
```

XGBoost将加载0002.model并进行两次迭代计算，并将输出明显保存在continue.model。需要注意的是 在mushroom.conf中定义的训练数据和评价数据信息不能发生变化。

使用多线程

当计算大数据集时，可能需要并行计算。如果编译器支持OpenMP，XGBoost原生是支持多线程的，通过一下参数 `nthread=10` 设置线程数为10。

其它需要注意的点

- `agaricus.txt.test.buffer` 和 `agaricus.txt.train.buffer` 是什么文件
 - 默认情况下XGBoost将会产生二进制的缓存文件，文件后缀为 `buffer`。当下次再次运行XGBoost时将加载缓存文件而不是原始的文件。



发表你的评论

(http://my.csdn.net/weixin_35068028)



touchphobia (/touchphobia) 2016-04-29 00:26

1楼

(/touchphobia) 请问这个文件`agaricus-lepiota.fmap`从哪来的？
Mushroom网页上好像没有这个

回复 1条回复

内容举报

返回顶部



工业设计公司




广告

相关文章推荐


XGBoost解决多分类问题 (<http://blog.csdn.net/u010159842/article/details/53411355>)

XGBoost解决多分类问题 写在前面的话 XGBoost官方给的二分类问题的例子是区别蘑菇有无毒，数据集和代码都可以在xgboost中的demo文件夹对应找到，我是用的Anaco...

 u010159842 (<http://blog.csdn.net/u010159842>) 2016年11月30日 17:58 2374

XGBoost：参数解释 (<http://blog.csdn.net/zc02051126/article/details/46711047>)

XGBoost参数在运行XGboost之前，必须设置三种类型成熟：general parameters，booster parameters和task parameters：General para...

 zc02051126 (<http://blog.csdn.net/zc02051126>) 2015年07月01日 17:06 49470




就刚刚，Python圈发生一件大事！

都说人生苦短，要学Python！但刚刚Python圈发生的这件事，你们怎么看？真相在这里...

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjnvPjn0IZ0qnfK9ujYzP1f4PjDs0Aw-5Hc3rHnYnHb0TAq15HfLPWRznjb0T1YLuHf4PjwWm10smH-BmHK90AwY5HDdnHckrHRznHn0lgF_5y9YIZ0IQzq-uZR8mLPbUB48ugfEIAqspynEmybz5LNYUNq1ULNzmvRqmhkEu1Ds0ZFb5HD0mhYqn0KsTWYs0ZNGujYkPHTYn1mk0AqGujYknWb3rjDY0APGujYLnWm4n1c0ULI85H00TZbqnW0v0APzm1YzP16vn0)


xgboost 二分类问题实例 (<http://blog.csdn.net/shenxiaoming77/article/details/76037930>)

二分类问题 本文介绍XGBoost的命令行使用方法。Python和R的使用方法见<https://github.com/dmlc/xgboost/blob/master/doc/README.md>。 ...

 shenxiaoming77 (<http://blog.csdn.net/shenxiaoming77>) 2017年07月24日 20:49 591


xgboost使用案例二 (<http://blog.csdn.net/hb707934728/article/details/70739382>)

-*- encoding:utf-8 -*- #xgboost安装教程 参考 http://blog.csdn.net/lht_okk/article/details/54311333 #xgbo...

 hb707934728 (<http://blog.csdn.net/hb707934728>) 2017年04月25日 14:49 1708

XGBoost解决多分类问题 (http://blog.csdn.net/Leo_Xu06/article/details/52424924)


内容举报


返回顶部




工业设计公司




内容举报

XGBoost解决多分类问题

 Leo_Xu06 (http://blog.csdn.net/Leo_Xu06) 2016年09月03日 19:25 66632

 返回顶部

一学就会的 WordPress 实战课



学习完本课程可以掌握基本的 WordPress 的开发能力，后续可以根据需要开发适合自己的主题、插件，打造最个性的 WordPress 站点。

(http://www.baidu.com/cb.php?c=lgF_pyfqHmknjfvP1m0IZ0qnfK9ujYzP1f4Pjnz0Aw-5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1YvnvwWuhm4Pj6zuHwBmvnL0AwY5HDdnHckrHRznHn0IgF_5y9YIZ0IQzqMpgwBUvq5HDknWw9mhkEusKzujYk0AFV5H00TZcqn0KdpyfqHRLPjnvnfKEpyfqHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45Hc)




Delphi7高级应用开发随书源码 (<http://download.csdn.net/detail/chensexh/3>)

(<http://download.csdn.net/detail/chensexh/3>) 2003年04月30日 00:00 676KB 下载


Xgboost的多分类 (<http://blog.csdn.net/wang1127248268/article/details/76769016>)

XGBoost解决多分类问题 XGBoost官方给的二分类问题的例子是区别蘑菇有无毒，数据集和代码都可以在xgboost中的demo文件夹对应找到，我是用的Anaconda安装的XGBo...

 wang1127248268 (<http://blog.csdn.net/wang1127248268>) 2017年08月06日 14:29 288


XGBoost：多分类问题 (<http://blog.csdn.net/zc02051126/article/details/46771243>)

下面用数据 UCI Dermatology dataset演示XGBoost的多分类问题首先要安装好XGBoost的C++版本和相应的Python模块，然后执行如下脚本，如果本地没有训练所需要的数据，...

 zc02051126 (<http://blog.csdn.net/zc02051126>) 2015年07月06日 10:09 9719

XGBoost：二分类问题 (http://blog.csdn.net/flyinghorse_2012/article/details/50533363)

本文介绍XGBoost的命令行使用方法。Python和R的使用方法见<https://github.com/dmlc/xgboost/blob/master/doc/README.md>。下面将介...

 flyinghorse_2012 (http://blog.csdn.net/flyinghorse_2012) 2016年01月17日 20:43 937

xgboost原理 (<http://blog.csdn.net/a819825294/article/details/51206410>)




ivnj7hnHPWnjFhPAD1Pyn4uW99ujqdlAdxTv

 内容举报


 返回顶部

文章内容可能会相对比较多，读者可以点击上方目录，直接阅读自己感兴趣的章节。1.序 距离上一次编辑将近10个月，幸得爱可可老师（微博）推荐，访问量陡增。最近毕业论文与xgboost相关，于是重新写一下...

 a819825294 (<http://blog.csdn.net/a819825294>) 2016年04月21日 10:15 087235


XGBoost：二分类问题 (http://blog.csdn.net/levy_cui/article/details/60877008)

二分类问题 本文介绍XGBoost的命令行使用方法。Python和R的使用方法见<https://github.com/dmlc/xgboost/blob/master/doc/README.md>。...

 levy_cui (http://blog.csdn.net/levy_cui) 2017年03月08日 17:37 0479


XGBoost-Python完全调参指南-参数解释篇 (<http://blog.csdn.net/wzmsltw/article/details/50...>)

关于XGBoost的参数，发现已经有比较完善的翻译了。故本文转载其内容，并作了一些修改与拓展。原文链接见：<http://blog.csdn.net/zc02051126/article/detail...>

 wzmsltw (<http://blog.csdn.net/wzmsltw>) 2016年03月27日 22:28 038724


XGBoost：在Python中使用XGBoost (<http://blog.csdn.net/zc02051126/article/details/4677...>)

在Python中使用XGBoost下面将介绍XGBoost的Python模块，内容如下： * 编译及导入Python模块 * 数据接口 * 参数设置 * 训练模型 * 提前终止程序 * ...

 zc02051126 (<http://blog.csdn.net/zc02051126>) 2015年07月06日 11:27 066435


机器学习xgboost实战—手写数字识别 (http://blog.csdn.net/Eddy_zheng/article/details/504...)

1、xgboost 安装安装问题这里就不再做赘述，可参考前面写的博文：http://blog.csdn.net/eddy_zheng/article/details/501845632、...

 Eddy_zheng (http://blog.csdn.net/Eddy_zheng) 2016年01月11日 12:13 012727

xgboost入门与实战（原理篇）(<http://blog.csdn.net/sb19931201/article/details/52557382>)

xgboost入门与实战（原理篇）前言：xgboost是大规模并行boosted tree的工具，它是目前最快最好的开源boosted tree工具包，比常见的工具包快10倍以上。在数据科学方面...

 sb19931201 (<http://blog.csdn.net/sb19931201>) 2016年09月16日 20:26 045279

利用随机森林、xgboost、logistic回归、预测泰坦尼克号是否生还的乘客的生存数据可视化<http://blog.csdn.net/...>





 内容举报

 返回顶部



利用随机森林,xgboost,logistic回归,预测泰坦尼克号上遇害乘客的获救概率 (http://blog.csdn....

数据示例：,PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked,Embarked_C,E...

 hb707934728 (http://blog.csdn.net/hb707934728) 2017年04月25日 14:29  3516



XGBoost分析总结 (http://blog.csdn.net/baidu_36316735/article/details/53780375)

认识XGBoost是在参加kaggle网站的机器学习比赛上接触到的。听一些过来者说用xgboost，分分钟上top10%。然而我用过之后发现并没有显著提升，一定是理解不够。要知道在2015年的时候29...

 baidu_36316735 (http://blog.csdn.net/baidu_36316735) 2016年12月21日 09:17  2710



机器学习（四）--- 从gbdt到xgboost (http://blog.csdn.net/china1000/article/details/511068...)

gbdt（又称Gradient Boost Decision Tree），是一种迭代的决策树算法，该算法由多个决策树组成。它最早见于yahoo，后被广泛应用在搜索排序、点击率预估上。xgboost是...

 china1000 (http://blog.csdn.net/china1000) 2016年04月09日 19:34  17330



xgboost的使用简析 (http://blog.csdn.net/John159151/article/details/45549143)

前言——记得在阿里mlib实习的时候，大家都是用mlib下的GBDT来train model的。但由于mlib不是开源的，所以在公司外是不能够使用。后来参加kaggle比赛的时候，认识到一个GD...

 John159151 (http://blog.csdn.net/John159151) 2015年05月07日 02:43  19493

xgboost使用步骤 (http://blog.csdn.net/vfgbv/article/details/72828385)

1. user_index, training_data, label = make_train_set(train_start_date, train_end_date, test_start...

 vfgbv (http://blog.csdn.net/vfgbv) 2017年06月01日 09:51  941



 内容举报

 TOP
返回顶部