

原

大数据竞赛平台——Kaggle 入门

2014年12月14日 21:34:01

wepon_

阅读数：156461

标签：

Kaggle

机器学习

数据挖掘

python

更多

版权声明：本文为博主原创文章，未经博主允许不得转载。 <https://blog.csdn.net/u012162613/article/details/41929171>

大数据竞赛平台——Kaggle 入门篇

这篇文章适合那些刚接触Kaggle、想尽快熟悉Kaggle并且独立完成一个竞赛项目的网友，对于已经在Kaggle上——过的网友来说，大时间阅读本文。本文分为两部分介绍Kaggle，第一部分简单介绍Kaggle，第二部分将展示解决一个竞赛项目的: 呈。如有错误，请




1、Kaggle简介

Kaggle是一个数据分析的竞赛平台，网址：<https://www.kaggle.com/>







企业或者研究者可以将数据、问题描述、期望的指标发布到Kaggle上，以竞赛的形式向广大的数据科学家征集解决方案，类似于KDD-CUP（国际知识发现和数据挖掘竞赛）。Kaggle上的参赛者将数据下载下来，分析数据，然后运用机器学习、数据挖掘等知识，建立算法模型，解决问题得出结果，最后将结果提交，如果提交的结果符合指标要求并且在参赛者中排名靠比赛丰厚的奖金。更多内容可以参阅：[大数据众包平台](#)








下面我以图文的形式介绍Kaggle：

进入Kaggle网站：

Active Competitions			
Featured		Helping Santa's Helpers Jingle bells, Santa tells ...	24 days 206 teams \$20,000
		Click-Through Rate Prediction Predict whether a mobile ad will be clicked	57 days 843 teams \$15,000
		BCI Challenge @ NER 2015 A spell on you if you cannot detect errors!	2 months 122 teams \$1,000

这是当前正在火热进行的有奖比赛，有冠军杯形状的是“Featured”，译为“号召”，召集数据科学高手去参赛。下面那个灰色的有试剂瓶是“Research”，奖金少一点。这两个类别的比赛是有奖竞赛，难度自然不小，作为入门者，应该先做练习赛：

101		Data Science London + Scikit-learn Scikit-learn is an open-source machine learning library for Python. Give it a try here!	17 days 149 teams Knowledge
		When bag of words meets bags of popcorn Use Google's Word2Vec for movie reviews	12 months 30 teams Knowledge
		Digit Recognizer Classify handwritten digits using the famous MNIST data	12 months 495 teams Knowledge
		Titanic: Machine Learning from Disaster Predict survival on the Titanic (with tutorials in Excel, Python, R, and an introduction to Random Forests)	12 months 2124 teams
		Facial Keypoints Detection Detect the location of keypoints on face images	12 months 38 teams Knowledge
		First steps with Julia Identify characters from Google Street View Pictures + tutorial with Julia	12 months 32 teams Knowledge

Playground		Sentiment Analysis on Movie Reviews Classify the sentiment of sentences from the Rot dataset	26 days 61 teams Knowledge
		Finding Elo Predict a chess player's FIDE Elo rating from one game	3 months 88 teams Knowledge
		Billion Word Imputation Find and impute missing words in the billion word corpus	4 months 59 teams Knowledge
		Forest Cover Type Prediction Use cartographic variables to classify forest categories	4 months 925 teams Knowledge
		Bike Sharing Demand Forecast use of a city bikeshare system	5 months 1591 teams Knowledge
		Random Acts of Pizza Predicting altruism through free pizza	5 months 285 teams Knowledge
		Poker Rule Induction Determine the rules of a card game from a small number of observations	5 months 10 teams Knowledge

左图的比赛是“101”，右图的是“Playground”，都是练习赛，适合入门。入门Kaggle最好的方法就是独立完成101和Playground这两个练习。本文的第二部分将选101中的“Digit Recognition”作为讲解。

点击进入赛题“Digit Recognition”：



Knowledge • 496 teams

Digit Recognizer

Wed 25 Jul 2012

Thu 31 Dec 2015 (12 months to go)

Dashboard

Home

Data

Make a submission

Information

Description

Evaluation

Rules

Tutorial

Forum

Leaderboard

Visualization

My Team

GitHub

My Submissions

Competition Details

Get the Data

Make a submission

Classify handwritten digits using the famous MNIST data

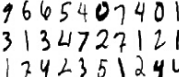
http://blog.csdn.net/u012162613

This competition is the first in a series of tutorial competitions designed to introduce people to Machine Learning.

The goal in this competition is to take an image of a handwritten single digit, and determine what that digit is. As the competition progresses, we will release tutorials which explain different machine learning algorithms and help you to get started.

The data for this competition were taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied. More detail about the dataset, including Machine Learning algorithms that have been tried on it and their results, can be found in the tutorial.

这是一个识别数字0~9的练习赛，“**Competition Details**”是这个比赛的描述，说明参赛者需要解决的问题。”**Get the Data**“是数据下载，用这些数据来训练自己的模型，得出结果，数据一般都是以csv格式给出：



Knowledge • 496 teams

Digit Recognizer

Wed 25 Jul 2012

Thu 31 Dec 2015 (12 months to go)

Dashboard

Home

Data

Make a submission

Information

Description

Evaluation

Rules

Tutorial

Forum

Leaderboard

Visualization

My Team

GitHub

My Submissions

Competition Details

Get the Data

Make a submission

Data Files

File Name	Available Formats
train	.csv (73.22 mb)
test	.csv (48.75 mb)
knn_benchmark	.R (316 b)
knn_benchmark	.csv (235.26 kb)
rf_benchmark	.R (381 b)
rf_benchmark	.csv (235.26 kb)

其中，train.csv就是训练样本，test.csv就是测试样本，由于这个是训练赛，所以还提供了两种解决方案，knn_benchmark.R和rf_benchmark.csv，前者是用R语言写的，后者是用Python写的。

得出结果后，接下来就是提交结果“**Make a submission**”：

要求提交的文件是csv格式的，假如你将结果保存在result.csv，那么点击“Click or drop submission here”，选中result.csv文件上传即可试你提交的结果的准确率，然后排名。

另外，除了“Competition Details”、“Get the Data”、“Make a submission”，侧边栏的“Home”、“Information”、“Forum”等，也提供了关些相关信息，包括排名、规则、辅导.....

【以上是第一部分，暂且写这么多，有补充的以后再更】

2、竞赛项目解题全过程

(1) 知识准备

首先，想解决上面的题目，还是需要一点ML算法的基础的，另外就是要会用编程语言和相应的第三方库来实现算法，常用的有：Python以及对应的库numpy、scipy、scikit-learn（实现了ML的一些算法，可以直接用）、theano（DeepLearning的算法包）。R语言、weka

如果用到深度学习的算法，cuda、caffe也可以用

总之，使用什么编程语言、什么平台、什么第三方库都无所谓，无论你用什麼方法，Kaggle只需要你线上提交结果，线下你如何实现限制的。

Ok，下面讲解题过程，以“Digit Recognition”为例，数字识别这个问题我之前写过两篇文章，分别用kNN算法和Logistic算法去实现，有码，有兴趣可以阅读：[kNN算法实现数字识别](#)、[Logistic回归实现数字识别](#)




(2) Digit Recognition解题过程

下面我将采用kNN算法来解决Kaggle上的这道Digit Recognition训练题。上面提到，我之前用kNN算法实现过，这里我将直接copy之前心代码，核心代码是关于kNN算法的主体实现，我不再赘述，我把重点放在处理数据上。

以下工程基于[Python、numpy](#)。

• 获取数据

从“Get the Data”下载以下三个csv文件：

 knn_benchmark.csv	2014/12/7 13:38	Microsoft Excel ...	236 KB
 test.csv	2014/12/7 14:13	Microsoft Excel ...	49,921 KB
 train.csv	2014/12/7 14:40	Microsoft Excel ...	74,976 KB

• 分析train.csv数据

train.csv是训练样本集，大小42001*785，第一行是文字描述，所以实际的样本数据大小是42000*785，其中第一列的每一个数字是它label，可以将第一列单独取出来，得到42000*1的向量trainLabel，剩下的就是42000*784的特征向量集trainData，所以从train.csv中取出两个矩阵trainLabel、trainData。

下面给出代码，另外关于如何从csv文件中读取数据，参阅：[csv模块的使用](#)

```

1 def loadTrainData():
2     l=[]
3     with open('train.csv') as file:
4         lines=csv.reader(file)
5         for line in lines:
6             l.append(line) #42001*785
7     l.remove(l[0])
8     l=array(l)
9     label=l[:,0]
10    data=l[:,1:]
11    return nomalizing(toInt(data)),toInt(label)

```

这里还有两个函数需要说明一下，toInt()函数，是将字符串转换为整数，因为从csv文件读取出来的，是字符串类型的，比如‘253’下来运算需要的是整数类型的，因此要转换，int(‘253’)=253。toInt()函数如下：

```

1 def toInt(array):
2     array=mat(array)
3     m,n=shape(array)
4     newArray=zeros((m,n))
5     for i in xrange(m):
6         for j in xrange(n):
7             newArray[i,j]=int(array[i,j])
8     return newArray

```

nomalizing()函数做的工作是归一化，因为train.csv里面提供的表示图像的数据是0~255的，为了简化运算，我们可以将其转化为二此将所有非0的数字，即1~255都归一化为1。nomalizing()函数如下：

```

1 def nomalizing(array):
2     m,n=shape(array)
3     for i in xrange(m):
4         for j in xrange(n):
5             if array[i,j]!=0:
6                 array[i,j]=1
7     return array

```

• 分析test.csv数据

test.csv里的数据大小是28001*784，第一行是文字描述，因此实际的测试数据样本是28000*784，与train.csv不同，没有label，28028000个测试样本，我们要做的工作就是为这28000个测试样本找出正确的label。所以从test.csv我们可以得到测试样本集testData，下：

```

1 def loadTestData():
2     l=[]
3     with open('test.csv') as file:
4         lines=csv.reader(file)
5         for line in lines:
6             l.append(line)
7     #28001*784
8     l.remove(l[0])
9     data=array(l)
10    return nomalizing(toInt(data))

```

• 分析knn_benchmark.csv

前面已经提到，由于digit recognition是训练赛，所以这个文件是官方给出的参考结果，本来可以不理这个文件的。但是我下面为了训练结果，所以也把knn_benchmark.csv这个文件读取出来，这个文件里的数据是28001*2，第一行是文字说明，去掉，第一列表2~28000，第二列是图片对应的数字。从knn_benchmark.csv可以得到28000*1的测试结果矩阵testResult，代码

```

1 def loadTestResult():
2     l=[]
3     with open('knn_benchmark.csv') as file:
4         lines=csv.reader(file)
5         for line in lines:
6             l.append(line)
7     #28001*2
8     l.remove(l[0])
9     label=array(l)
10    return toInt(label[:,1])

```

到这里，数据分析和处理已经完成，我们获得的矩阵有：trainData、trainLabel、testData、testResult

• 算法设计

这里我们采用KNN算法来分类，核心代码：

```

1 def classify(inX, dataSet, labels, k):
2     inX=mat(inX)
3     dataSet=mat(dataSet)
4     labels=mat(labels)
5     dataSetSize = dataSet.shape[0]
6     diffMat = tile(inX, (dataSetSize,1)) - dataSet
7     sqDiffMat = array(diffMat)**2
8     sqDistances = sqDiffMat.sum(axis=1)
9     distances = sqDistances**0.5
10    sortedDistIndicies = distances.argsort()
11    classCount={}
12    for i in range(k):
13        voteIlabel = labels[0,sortedDistIndicies[i]]
14        classCount[voteIlabel] = classCount.get(voteIlabel,0) + 1
15    sortedClassCount = sorted(classCount.iteritems(), key=operator.itemgetter(1), reverse=True)
16    return sortedClassCount[0][0]

```

关于这个函数，参考：[kNN算法实现数字识别](#)

简单说明一下，inX就是输入的单个样本，是一个特征向量。dataSet是训练样本，对应上面的trainData，labels对应trainLabel，k指定的k，一般选择0~20之间的数字。这个函数将返回inX的label，即图片inX对应的数字。

对于测试集里28000个样本，调用28000次这个函数即可。

• 保存结果

kaggle上要求提交的文件格式是csv，上面我们得到了28000个测试样本的label，必须将其保存成csv格式文件才可以提交，关于csv，请参考[【Python】csv模块的使用](#)。

代码:

```
1 def saveResult(result):
2     with open('result.csv','wb') as myFile:
3         myWriter=csv.writer(myFile)
4         for i in result:
5             tmp=[]
6             tmp.append(i)
7             myWriter.writerow(tmp)
```

• 综合各函数

上面各个函数已经做完了所有需要做的工作，现在需要写一个函数将它们组合起来解决digit recognition这个题目。我们写一个handwritingClassTest函数，运行这个函数，就可以得到训练结果result.csv。

```
1 def handwritingClassTest():
2     trainData,trainLabel=loadTrainData()
3     testData=loadTestData()
4     testLabel=loadTestResult()
5     m,n=shape(testData)
6     errorCount=0
7     resultList=[]
8     for i in range(m):
9         classifierResult = classify(testData[i], trainData, trainLabel, 5)
10        resultList.append(classifierResult)
11        print "the classifier came back with: %d, the real answer is: %d" % (classifierResult, testLabel[0,i])
12        if (classifierResult != testLabel[0,i]): errorCount += 1.0
13    print "\nthe total number of errors is: %d" % errorCount
14    print "\nthe total error rate is: %f" % (errorCount/float(m))
15    saveResult(resultList)
```

运行这个函数，可以得到result.csv文件：

	A	B
1	2	
2	0	
3	9	
4	9	
5	3	
6	7	
7	0	
8	3	
9	0	
10	3	
11	5	
12	7	
13	4	
14	0	
15	4	
16	5	
17	3	
18	1	
19	9	
20	0	
21	9	
22	1	
23	1	
24	5	
25	7	
26	4	
27	2	
--	--	

2 0 9 9 3 7 0 3.....就是每个图片对应的数字。与参考结果knn_benchmark.csv比较一下：

```

the classifier came back with: 2, the real answer is: 2
the classifier came back with: 2, the real answer is: 2
the classifier came back with: 9, the real answer is: 9
the classifier came back with: 6, the real answer is: 6
the classifier came back with: 7, the real answer is: 7
the classifier came back with: 6, the real answer is: 6
the classifier came back with: 1, the real answer is: 1
the classifier came back with: 9, the real answer is: 9
the classifier came back with: 7, the real answer is: 7
the classifier came back with: 9, the real answer is: 9
the classifier came back with: 7, the real answer is: 7
the classifier came back with: 7, the real answer is: 7
the classifier came back with: 9, the real answer is: 9
the classifier came back with: 7, the real answer is: 7
the classifier came back with: 3, the real answer is: 3
the classifier came back with: 9, the real answer is: 9
the classifier came back with: 2, the real answer is: 2

the total number of errors is: 1004
the total error rate is: 0.035857

```

28000个样本中有1004个与kkn_benchmark.csv中的不一样。错误率为3.5%，这个效果并不好，原因是我并未将所有训练样本都拿来训练时间，我只取一半的训练样本来训练，即上面的结果对应的代码是：

```
classifierResult = classify(testData[i], trainData[0:20000], trainLabel[0:20000], 5)
```

训练一半的样本，程序跑了将近70分钟（在个人PC上）。

• 提交结果

将result.csv整理成kkn_benchmark.csv那种格式，即加入第一行文字说明，加入第一列的图片序号，然后make a submission，结果96.5%：

314	↓30	Steve Shank	0.96557	2	Fri, 05 Dec 2014 18:34:04 (-0.8h)
315	↓30	raito	0.96557	4	Sun, 07 Dec 2014 03:49:52 (-11.6h)
316	↓30	wepon	0.96557	3	Sun, 14 Dec 2014 14:34:44 (-7.4d)
317	new	chiwei	0.96557	1	Tue, 09 Dec 2014 07:06:53
318	new	Stan Valchek	0.96557	1	Thu, 11 Dec 2014 18:54:04

下载工程代码：[github地址](#)

【完】

Python爬虫全栈教学，零基础教你成编程大神
零基础学爬虫，你要掌握学习那些技能？

想对作者说点什么？

我来说两句

**床长**：写得不错！我最近也在写一系列的人工智能教程，通俗易懂，无需高等数学基础,教程也力求风趣幽默。点击我的头像浏览教程，最好从序言看起。希望更多的工智能大家庭中，使中国更加强大，使中国人在国外能把头抬的更高！（8个月前 #28楼）

**kovoja**：感谢分享，算是入门介绍了！（9个月前 #27楼）

**Anthony_azy**：感谢分享，我使用了全部数据进行测试，准确率才96.4%（11个月前 #26楼）[查看回复\(1\)](#)

查看 47 条热评

Kaggle竞赛入门教程之Kaggle简介（新手向）
Kaggle是全球最大的数据科学家汇聚的平台，机器学习高手云集，同时对萌新也很友好。Kaggle网址：<https://www.kaggle.com/>

2.5万

来自：[大家好，我是Utanbo](#)

【Kaggle】Titanic详解
kaggle：<https://www.kaggle.com/c/titanic>这里做一个简单笔记记录提交准确率：0.83代码详解:1、数据读取#读取训...

280

**有哪些可以免费试用一年左右的云服务器**
百度广告

关于Kaggle入门，看这一篇就够了
这次酝酿了很久想给大家讲一些关于Kaggle那点事儿，帮助对数据科学(Data Science)有兴趣的同学们更好的了解...

1.7万

来自：[bbbeoy的专栏](#)

Kaggle 机器学习竞赛冠军及优胜者的源代码汇总
阅读目录 Algorithmic Trading Challenge25Allstate Purchase Prediction Challenge3Amazon.com – Employee ...

1.5万

来自：[Leo的博客](#)

Kaggle初学者五步入门指南，七大诀窍助你享受竞赛
Kaggle初学者五步入门指南，七大诀窍助你享受竞赛 By 机器之心2017年7月22日 14:41 Kaggle 是一个流行的数据...

2.7万

来自：[Slark的博客](#)

Kaggle 新手教程（一）
在DATAQUEST上学习kaggle的教程，感觉有些数据预处理的代码很实用，并且用的是之前没接触过的pandas写的...

2049

来自：[isaacfeng的博客](#)

光谷股王8年追涨停铁律“1272”曝光，震惊众人
第六·熾燚

Kaggle入门
由于选修了数据挖掘课程，课程作业是完成Kaggle上的一个比赛，所以在机缘巧合下就知道了Kaggle这个平台，事...

1.6万

来自：[acelove40的博客](#)

Kaggle 首战拿银总结 | 入门指导 (长文、干货)
Kaggle 首战拿银总结 | 入门指导 (长文、干货)

1.5万

来自：[技术博客](#)

https://blog.csdn.net/u012162613/article/details/41929171?utm_source=blogxgwz4

8/11

Scikit-learn快速入门教程和实例（一）

Github主页：https://linxid.github.io/ 知乎：https://www.zhihu.com/people/dong-wen-hui-90/activities ...

8561

来自：linxid的博客

【机器学习算法实现】kNN算法_手写识别——基于Python和NumPy函数库

kNN算法，即K最近邻(k-NearestNeighbor)分类算法，是最简单的机器学习算法之一，算法思想很简单：从训练样本...

2.7万

来自：wepon的专栏

相关热词 大数据前端 大数据及计算方法 大数据发展史 大数据存储引擎 大数据学习网

博主推荐



天涯泪小武
[关注](#) 188篇文章



李博Garvin
[关注](#) 290篇文章



oxuzhenyi
[关注](#) 183篇文章



wepon_
[关注](#)

原创	粉丝	喜欢	评论
72	2911	180	491

等级： **博客 5** 访问：136万+

积分：6405 排名：5493

勋章： 




可视化大数据



About

个人网站：http://wepon.me/
github：https://github.com/wepe
知乎：
https://www.zhihu.com/people/wepon-huang
很久没上CSDN，有问题欢迎邮件交流：
masterwepon@163.com

博主专栏



深度学习入门指南
阅读量：207730 5 篇

个人分类

Machine Learning	31篇
Kaggle	2篇
scikit-learn使用手册	1篇
python	16篇
数据库	1篇

展开

归档	
2016年2月	1篇
2015年9月	1篇
2015年8月	1篇
2015年5月	3篇
2015年4月	3篇
展开	

热门文章

大数据竞赛平台——Kaggle 入门

阅读量：156364

交叉熵代价函数

阅读量：121850

正则化方法：L1和L2 regularization、数据集扩增、dropout

阅读量：73978

DeepLearning tutorial (5) CNN卷积神经网络应用于人脸识别（详细流程+代码实现）

阅读量：73351

DeepLearning tutorial (4) CNN卷积神经网络原理简介+代码详解

阅读量：68821

最新评论

DeepLearning tuto...
yuan0401yu：[reply]qq_40975575[/reply] 我改成了pool_2d 就对了

【机器学习算法实现】主成分分析(P...
weixin_43332451：[reply]weixin_38364427[/reply] 楼主的代码是对的，你的想法也对。但是...

【csapp】【微软面试题】有符号...
essity：第三个问题中，是将unsigned int转换成int吧，这样的话，就不是恒大于0了。

【机器学习算法实现】kNN算法____...
qq_40678277：[reply]jiangjunshow[/reply] 你写的太垃圾了

DeepLearning tuto...
u014483682：[reply]kobecsb[/reply] 18年怎么还会有人这样想

loft公寓出租



联系我们



扫码联系客服



官方公众号


 kefu@csdn.net

 400-660-0108

 QQ客服

 客服论坛

[关于我们](#)[招聘](#)[广告服务](#)[网站地图](#)

 百度提供站内搜索 京ICP证09002463号

©2018 CSDN版权所有

[经营性网站备案信息](#)[网络110报警服务](#)

[北京互联网违法和不良信息举报中心](#)

[中国互联网举报中心](#)