

Coggle 30 Days of ML (21年12月)

Part1 内容介绍

在给大家分享知识的过程中，发现很多同学在学习竞赛都存在较多的问题：

- 不知道如何使用Pandas查询数据、分析数据
- 不知道如何入手结构化比赛
- 不知道如何搭建模型

而上述问题都是一个竞赛选手、一个算法工程师所必备的。因此我们将从本月组织一次竞赛训练营活动，希望能够帮助大家入门数据竞赛。在活动中我们将布置具体竞赛任务，然后参与的小伙伴们不断闯关完成，尽可能的帮助大家入门。

Part2 活动安排

- 活动是免费学习活动，不会收取任何费用。
- 请各位同学添加下面微信，并回复【Pandas、反欺诈】，即可参与。



Part3 积分说明和奖励

为了激励各位同学完成的学习任务，将学习任务根据难度进行划分，并根据是否完成进行评分难度高中低的任务分别分数为 3、2 和 1。在完成 12月学习后，将按照积分顺序进行评选 Top3 的学习者。

积分排名更新时间：12月28日

打卡链接：<https://shimo.im/forms/axVsWeunl3QEjGWy/fill>

积分有问题可以联系小助手哦！

打卡可以写在一个地址，每次有新完成的可以重复提交打卡！

微信昵称	Pandas数据处理	广告反欺诈比赛	总积分
simula67	4	1	5
胡子大叔	Notebook没有设置公开		
Freddy	Notebook没有设置公开	Notebook没有设置公开	
今天吃到七分饱就收手了吗			
ysy			
切追风			
他说民谣很穷			
回想			
小六子			
HITSZ-白			
香蕉拌豆腐			
快快长高			
御史大浩浩	21	8	29
36V	1		1
Echo			
xuxiaoshunweixin	4		4
QDD			
无盐	4		4
大西几			
melondeath	11	3	14
无盐	4	3	7
Giv Xs茯苓ωの茯苓&梦	18		18
Keesh		3	3
Peter	4		4
wzs	21		21
初心	4	2	6
Ceallach_Shaw	1		1
zzz963421397	2	Notebook没有设置公开	2
糖醋鱼	21	20	41

slowwalkor			
哈哈😄	21	1	22
吴定俊	11	3	14
小米粥铺	4		4
我的肚子圆鼓鼓	4		4
z	4	Notebook没有设置公开	4
chen	4		4
PURPLE	21		21
zhangxue	21	3	24
无盐	21	3	24
Freddy	Notebook没有设置公开		
黄文川	4		4
Cyan	21		21
jarvis1890	21		21
PURPLE	18		18
李宽	21		21
LPF	21		21
飞羽	11	20	31
大西几	15		15
徐乜乜	21	20	41
Ceallach_Shaw	18	10	28
柴门犬吠	13		15
ys		3	3
宁静致远	9	0	9
方塘 Ricky	18		18
正牌可口可乐	18		18
Young Cat		8	8

Top1的学习者将获得以下**奖励**：

- Coggle 竞赛专访机会
- 《机器学习算法竞赛实战》，鱼佬签名版

Top2–3的学习者将获得以下**奖励**：

- Coggle 周边福利
- Coggle 竞赛专访机会

使用PaddlePaddle完成学习的Top3同学，还可以领取百度提供的小礼物。

注：

- Coggle 数据科学保留活动期间和结束后修改奖励和规则的权利。
- 如果有违反竞赛规则的情况，Coggle 数据科学保留取消相关参赛者的参与排名的权利。

Part4 Pandas数据处理

学习内容

Pandas 是一个开放源码、BSD 许可的库，提供高性能、易于使用的数据结构和数据分析工具。Pandas 名字衍生自术语 "panel data"（面板数据）和 "Python data analysis"（Python 数据分析）。

Pandas 一个强大的分析结构化数据的工具集，基础是 Numpy（提供高性能的矩阵运算）。在 Python 环境下，Pandas 是数据挖掘、机器学习和深度学习必备的基础库，可以极大的提高数据处理效率。

打卡汇总

任务名称	难度	所需技能
任务1：Pandas数据读取、保存和数据类型	低、1	DataFrame
任务2：Pandas数据位置索引	低、1	loc、iloc
任务3：Pandas数据逻辑索引	中、2	loc、where
任务4：Pandas数据分组聚合	高、3	groupby、agg、transform
任务5：Pandas日期数据处理	中、2	dt
任务6：Pandas缺失值处理	中、2	fillna
任务7：Pandas数据可视化	中、2	plot
任务8：Pandas多表合并和聚合	中、2	merge
任务9：Pandas透视表和交叉表操作	高、3	
任务10：Pandas性能优化	高、3	Numpy、joblib

学习资料

<https://github.com/datawhalechina/joyful-pandas>

<https://blog.csdn.net/u010161379/article/details/79187614>

打卡要求

注：

- 需要所有的任务可以写在一个Notebook内
- 推荐在打卡过程中加入思考过程，可以加入尝试&资料记录

- 打卡Notebook必须在百度 AI Studio平台运行，并设置公开

任务1：Pandas数据读取、保存和数据类型

任务要点：文件读取、保存、数据类型分析

- 步骤1：读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2：将读取的进行保存，表头也需要保存
- 步骤3：分析每列的类型，取值个数
- 步骤4：分析每列是否包含缺失值

任务2：Pandas数据位置索引

任务要点：数据选择、数据索引

- 步骤1：读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2：选择出Total列
- 步骤3：选择出Total列和HP列
- 步骤4：选择出第10-40行数据
- 步骤5：选择出第10-40行的Total列和HP列

任务3：Pandas数据逻辑索引

任务要点：逻辑索引

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.where.html>

https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html

- 步骤1：读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2：筛选出Type 1 为 Grass的数据
- 步骤3：筛选出Type 1 为 Grass的数据 且 Type 2 为 Poison的数据
- 步骤4：筛选出HP大于50 或 Speed小于90的数据
- 步骤5：筛选出Type 1 取值为Grass 或 Fire，且 HP 位于 70 与 90之间，且 Speed以数字8 开头的的数据

任务4：Pandas数据分组聚合

https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html

任务要点：groupby、agg、transform

- 步骤1：读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2：学习groupby分组聚合的使用
- 步骤3：学习agg分组聚合的使用
- 步骤4：学习transform的使用
- 步骤5：使用groupby、agg、transform，统计数据在Type 1分组下 HP的均值

任务5：Pandas日期数据处理

https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html

https://pandas.pydata.org/pandas-docs/stable/user_guide/timedeltas.html

任务要点：日期处理

- 步骤1：创建一列dt，dt取值为从1638263656 到 1638283656 的 unix时间
- 步骤2：将dt列转为datetime格式
- 步骤3：筛选出dt列中小时为10的行
- 步骤4：将dt列整体增加8小时的时间

任务6: Pandas缺失值处理

https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

- 步骤1: 读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2: 分析每列的缺失值
- 步骤3: 对每列的缺失值进行填空

任务7: Pandas数据可视化

<https://www.cnblogs.com/zhangyafei/p/10518826.html>

https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html

任务要点: plot

- 步骤1: 读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2: 统计Type 1分组下HP、Attack、Defense的均值, 并进行绘制柱状图
- 步骤3: 将所有样本的HP、Attack绘制散点图

任务8: Pandas多表合并和聚合

<https://blog.csdn.net/jasonzhoujx/article/details/81665558>

https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html

<https://queirozf.com/entries/pandas-dataframes-merge-join-examples>

ID X1			ID X2		
1	a1		2	b1	
2	a2		3	b2	

Inner Join			Outer Join			Left Join			Right Join		
ID	X1	X2	ID	X1	X2	ID	X1	X2	ID	X1	X2
2	a2	b1	1	a1	NA	1	a1	NA	2	a2	b1
			2	a2	b1	2	a2	b1	3	NA	b2
			3	NA	b2						

任务要点: merge、join

- 步骤1: 创建如下数据

```
1 data1 = pd.DataFrame({"ID":range(101, 106),
2                       "x1":range(1, 6),
3                       "x2":["a", "b", "c", "d", "e"],
4                       "x3":range(16, 11, - 1)})
5
6 data2 = pd.DataFrame({"ID":range(104, 108),
7                       "y1":["x", "y", "x", "y"],
8                       "y2":range(8, 1, - 2)})
9
10 data3 = pd.DataFrame({"ID":range(102, 110),
11                       "z1":range(10, 18),
12                       "z2":["z", "b", "z", "z", "d", "z", "d", "a"],
```

```
"z3":range(18, 10, - 1)}))
```

- 步骤2: Merge data1 and data2 Using Inner Join
- 步骤3: Merge data1 and data2 Using Outer Join
- 步骤4: Merge data1 and data2 Using Left Join
- 步骤5: Merge data1 and data2 Using Right Join
- 步骤6: Merge data1 ,data2 and data3 Using Outer Join
- 步骤7: Merge data1 and data2 based on Index

任务9: Pandas透视表和交叉表操作

https://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html

Pivot

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

➔

```
df.pivot(index='foo',
          columns='bar',
          values='baz')
```

	bar	A	B	C
foo				
one		1	2	3
two		4	5	6

- 步骤1: 读取文件<https://cdn.coggle.club/Pokemon.csv>
- 步骤2: 统计Type 1和Type 2分组下HP,Attack,Defense的均值
- 步骤3: 统计Type 1为index, Type 2取值为不同列的情况下, Attack的均值

任务10: Pandas性能优化

https://pandas.pydata.org/pandas-docs/stable/user_guide/enhancingperf.html

<https://realpython.com/fast-flexible-pandas/>

- 步骤1: 创建如下数据

```
1 data = pd.DataFrame({
2     "x1":list(range(1000, 2000))*1000,
3     "x2":list(range(1000))*1000
4 })
```

- 步骤2: 使用iloc遍历数据集, 并记录下时间。
- 步骤3: 使用 `itertuples` 和 `iterrows` 遍历数据集, 并记录下时间。
- 步骤4: 使用cython和ndarray完成数据遍历, 并统计统计下x1分组下x2的均值。

Part5 广告反欺诈比赛

赛题介绍



常规赛：MarTech Challenge 点击反欺诈预测 进行中

提供约50万次点击数据，请预测用户的点击行为是否为正常点击，还是作弊行为

标签：反欺诈预测 比赛时间：2020/09/01 - 2023/01/01 主办方：Baidu 大脑

[立即报名](#)

<https://aistudio.baidu.com/aistudio/competition/detail/52/0/introduction>

广告欺诈是数字营销需要面临的重要挑战之一，点击会欺诈浪费广告主大量金钱，同时对点击数据会产生误导作用。本次比赛提供了约50万次点击数据。特别注意：我们对数据进行了模拟生成，对某些特征含义进行了隐藏，并进行了脱敏处理。

请预测用户的点击行为是否为正常点击，还是作弊行为。点击欺诈预测适用于各种信息流广告投放，banner广告投放，以及百度网盟平台，帮助商家鉴别点击欺诈，锁定精准真实用户。

字段	类型	说明
sid	string	样本id/请求会话sid
package	string	媒体信息，包名（已加密）
version	string	媒体信息，app版本
android_id	string	媒体信息，对外广告位ID（已加密）
media_id	string	媒体信息，对外媒体ID（已加密）
apptype	int	媒体信息，app所属分类
timestamp	bigint	请求到达服务时间，单位ms
location	int	用户地理位置编码（精确到城市）
fea_hash	int	用户特征编码（具体物理含义略去）
fea1_hash	int	用户特征编码（具体物理含义略去）
cus_type	int	用户特征编码（具体物理含义略去）
ntt	int	网络类型 0-未知, 1-有线网, 2-WIFI, 3-蜂窝 未知, 4-2G, 5-3G, 6-4G
carrier	string	设备使用的运营商 0-未知, 46000-移动, 46001-联通, 46003-电信
os	string	操作系统，默认为android
osv	string	操作系统版本
lan	string	设备采用的语言，默认为中文
dev_height	int	设备高
dev_width	int	设备宽
dev_ppi	int	屏幕分辨率

打卡汇总

任务名称	难度	所需技能
任务1：报名比赛，下载比赛数据集并完成读取	低、1	Pandas
任务2：对数据字段理解，对特征字段依次进行数据分析	中、2	Matplotlib/Seaborn
任务3：使用特征工程对比赛字段进行编码	高、3	Pandas、Sklearn
任务4：使用 Sklearn 中基础树模型完成训练和预测提交	中、2	Sklearn
任务5：使用 Sklearn 中线性模型完成训练和预测提交	低、1	Sklearn
任务6：使用特征重要性分析方法分析特征重要性	中、2	shap
【Paddle学习部分】		
任务7：使用Paddle完成基础MLP完成训练和预测提交	中、2	PaddlePaddle
任务8：使用Paddle完成类别的Embedding处理，搭建Wide & Deep训练和预测提交	低、1	PaddlePaddle
任务9：学习FM、FFM的原理和基础实现	高、3	PaddlePaddle
任务10：使用Paddle搭建DeepFM模型完成训练和预测提交	高、3	PaddlePaddle
【非Paddle学习部分】		
任务7：完成基础MLP完成训练和预测提交	中、2	Pytorch、Keras
任务8：完成类别的Embedding处理，搭建Wide & Deep模型完成训练和预测提交	低、1	Pytorch、Keras
任务9：学习FM、FFM的原理和基础实现	高、3	Pytorch、Keras
任务10：搭建DeepFM模型完成训练和预测提交	高、3	Pytorch、Keras

打卡要求

注：

- 需要所有的任务可以写在一个Notebook内
- 推荐在打卡过程中加入思考过程，可以加入尝试&资料记录
- 若使用Paddle进行打卡必须在百度 AI Studio平台运行，并设置公开

任务1：报名比赛，下载比赛数据集并完成读取

任务要点：数据读取

- 步骤1：报名比赛，并下载比赛数据集<https://aistudio.baidu.com/aistudio/competition/detail/52/0/introduction>
- 步骤2：使用Pandas完成数据读取

任务2：对数据字段进行理解，对特征字段依次进行数据分析

任务要点：数据分析、数据统计

- 步骤1：数据字段取值分析，每个字段的取值范围、类型
- 步骤2：数据字段分布分析，每个字段的整体分布
- 步骤3：数据字段的相关性分析
- 步骤4（可选）：字段与标签的EDA探索性分析

任务3：使用特征工程对比赛字段进行编码



Feature+Engineering.pdf
0.9MB

预览

```
1 from sklearn.preprocessing import OneHotEncoder
2 enc = OneHotEncoder(handle_unknown='ignore')
3 X = [['Male', 1], ['Female', 3], ['Female', 2]]
4
5 enc.fit(X)
6 enc.transform(['Female', 1], ['Male', 4]).toarray()
```

```
1 import pandas as pd
2 df = pd.DataFrame({'A': ['a', 'b', 'a'], 'B': ['b', 'a', 'c'],
3                  'C': [1, 2, 3]})
4
5 pd.get_dummies(df, prefix=['col1', 'col2'])
```

任务4：使用 Sklearn 中基础树模型完成训练和预测

☐ 学会五折交叉验证的数据划分方法（KFold）

```
1 import numpy as np
2 from sklearn.model_selection import KFold
3 X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
4 y = np.array([1, 2, 3, 4])
5 kf = KFold(n_splits=2)
6
7 for train_index, test_index in kf.split(X):
8     print("TRAIN:", train_index, "TEST:", test_index)
9     X_train, X_test = X[train_index], X[test_index]
10    y_train, y_test = y[train_index], y[test_index]
```

任务5：使用 Sklearn 中线性模型完成训练和预测

任务6：使用树模型和线性模型，分析特征重要性

- ☐ 步骤1：树模型的特征重要性计算
- ☐ 步骤1：线性模型的特征重要性计算

任务7：使用基础MLP完成训练和预测提交

- ☐ 步骤1：使用PaddlePaddle或Pytorch搭建基础的全连接网络
- ☐ 步骤2：训练全连接网络并提交

任务8：完成类别的Embedding处理

- ☐ 步骤1：使用PaddlePaddle或Pytorch搭建完成类别的Embedding嵌入操作
- ☐ 步骤2：将Embedding搭配全连接网络，搭建Wide & Deep模型完成训练和预测提交

任务9：学习FM、FFM的原理和基础实现

- ☐ 步骤1：学习FM原理, https://blog.csdn.net/qq_27782503/article/details/109069750
- ☐ 步骤2：学习FFM原理, <https://blog.csdn.net/leisurehippo/article/details/74226111>

任务10：搭建DeepFM模型完成训练和预测提交

- ☐ 步骤1：使用PaddlePaddle或Pytorch搭建DeepFM模型, <https://zhuanlan.zhihu.com/p/91151350>
 - ☐ 步骤2：DeepFM模型训练与提交
-

Part6 提问&回答

问：具体的活动是怎么安排的？

有任务，自己先尝试。活动结束后会公开优秀打卡链接。

问：本次活动是收费的吗，最终奖品如何发放？

活动是免费的，最终奖品按照积分排行Top3进行发放，如果排名有并列都发送奖励。

问：环境和配置是什么？

在AI Studio上进行学习，python3和PaddlePaddle环境

问：AI Studio有什么学习资料？

项目环境介绍: <https://ai.baidu.com/ai-doc/AISTUDIO/Dk3e2vxg9>
Notebook环境: <https://ai.baidu.com/ai-doc/AISTUDIO/sk3e2z8sb>