第十五节课 CTR特征组合与机器学习开发工具

https://github.com/cystanford 老师的GitHub

第十五节课笔记

讲义内容概括	知识点	Question?

CTR中的特征组合：
对于机器学习来说，特征选择的重要性毋庸置疑 => 如何进行特征选择、高效的组合特征
对于推荐系统来说，DeepFM, Wide&Deep帮我们解决了一部分特征组合的问题，这些模型可以高效的进行特征交互，将原本的低维稀疏特征转换到高维稠密特征，再用NN进行处理，效率和效果都有质的提升 DCN, xDeepFM对原有模型进行改进，提出新的特征组合方式，提升了效果

针对数据量大的情况：
数据存储和传输：OSS
数据计算：NoteBook
下采样：行采样，列采样 => 小样本数据集
对avazu数据集进行下采样，生成小样本后传回给OSS

chunksize使用：

- pandas使用chunksize分块处理大型csv文件
- chunksize，单个IO大小，设置越大站占用内存高，需要的iteration少，速度快

```
file_path = '/home/admin/avazu/'

df = pd.read_csv(file_path + 'sub_train_f.csv',
chunksize=1000)

for chunk in df:

    print(chunk)

    print('-'*1000)
```

CTR特征组合
Project A：Avazu CTR广告点击率预测
Project B：天猫用户复购预测
CTR预估
Review MF, FM Deep & Cross模型
xDeepFM模型
如何对大数据集进行处理

机器学习开发工具
特征组合工具FeatureTools
使用ImageAI本地完成AI应用
Serverless架构下的AI应用
腾讯广告算法大赛2019
腾讯广告算法大赛2020
Baidu Track2 2020

TIPS：
规律产生价值
各种模型 => 规律
使用规律 or 创造规律
对于机器学习优化方法

gc使用：

- Garbage Collection，各大语言对数据处理的必备工具之一
- 本质是内存的自动管理，用来回收堆（Heap）中不再需要（使用）的对象

```
# 一个chunk块为5万行
chunksize = 50000
df_train_f = pd.read_csv(fp_train_f, dtype={'id':str},
index_col=None, chunksize=chunksize, iterator=True)

import gc
del lr_model
del df_train_f
gc.collect()
```

大数据处理（pkl）

存储临时文件（变量）

- pkl文件

```
df = pd.read_pickle("oh_enc.pkl")
result.to_pickle("oh_enc.pkl")
```

处理好的数据存到硬盘里，存成pkl格式

下次读取的时候加快读取速度

- 使用pickle

用于python特有的类型和python的数据类型间进行转换

- 四个功能：dumps,dump,loads,load

```
import pickle
obj = 123, "abcdef", ["ac", 123], {"key": "value", "key1": "value1"}
print(obj)
# 序列化到文件
file_path = '/home/admin/temp/'
pickle.dump(obj, open(file_path + 'a.txt', 'wb'))
temp = pickle.load(open(file_path + 'a.txt', 'rb'))
print(temp)
```

大数据处理（partial_fit）

partial_fit 递增式学习

- partial_fit(X,y,classes=None,sample_weight=None)

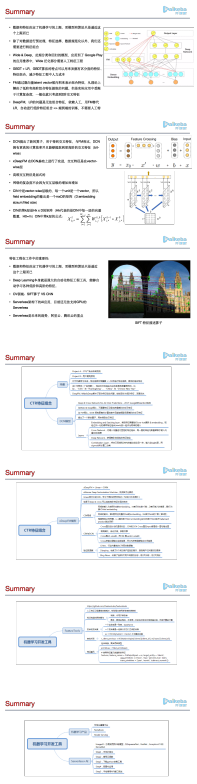
参数

- X:样本数据
- y:样本标记
- classes:列出所有可能的类别
- sample_weight:给出每个样本的权重(未指定，则全为1)

partial_fit方法可以使得几十GB的数据集被切分成一块一块的来递增训练，每次最好数据块足够大（充分利用内存）

```
from sklearn.linear_model import SGDClassifier
clf = SGDClassifier(loss='log')
#用数据集训练
clf.partial_fit(X, y)
#当我们有了新数据之后，可以在原基础上更新模型
clf.partial_fit(X_new, y_new)
#partial_fit的模型使用方法也是和正常模型一样的，直接用predict或者predict_proba
y_pred = clf.predict_proba(X_test)
```

结果来说，特征选择可能比模型选择更重要
各种新模型提出，首先都是对原有特征规律的新洞察



Python算法思考题:
句子相似度并查集

- SGD，算法收敛速度快，但容易收敛到局部最优。SGD的缺点是更新方向依赖于当前batch计算出的梯度，因而很不稳定
- Momentum，借用了物理中的动量概念，更新的时候在一定程度上保留之前更新的方向，同时利用当前batch的梯度微调最终的更新方向，即模拟了运动惯性。
- Adagrad，自适应梯度算法，能够在训练中自动的对learning rate进行调整，对于出现频率较低参数采用较大的α更新，对于出现频率较高的参数采用较小的α更新
- RMSprop，均方根传播，Adagrad会累加之前所有的梯度平方，而RMSprop仅仅是计算对应的平均值，可以缓解Adagrad算法学习率下降较快的问题
- Adam， 结合了 AdaGrad 和 RMSProp 算法最优的性能，它还是能提供解决稀疏梯度和噪声问题的优化方法，在深度学习中使用较多

人工特征工程
转换，作用于单张表，这些转换操作都只用到了一张表的信息
聚合，跨表实现的，并使用一对多的关联来对观测值分组，然后计算统计量
聚合，跨表实现的比如有一张客户贷款表，每个客户可能有多项贷款，可以计算每个客户贷款的平均值、最大值和最小值等统计量
表的关联：创建关联并将其添加到实体集中的语法：
FeatureTools工具

```
# 创建表之间的关联
r_client_previous = ft.Relationship(es['clients']['client_id'],
es['loans']['client_id'])
r_payments = ft.Relationship(es['loans']['loan_id'],
es['payments']['loan_id'])
# 添加关系到实体集中
es = es.add_relationship(r_client_previous)
es = es.add_relationship(r_payments)
print(es)
```

特征基元：
转换：对一张表中一或多列完成的操作。比如取一张表中两列之间的差值，或者取一列的绝对值
聚合：根据父与子（一对多）的关联完成的操作，也就是根据

父亲分组并计算儿子的统计量。比如根据 client_id 对 loan 表分组，并找到每个客户的最大贷款额
https://github.com/FeatureLabs/featuretools

Serverless架构下的AI应用
部署一个Serverless的图像识别应用：
Step1，本地训练AI
ImageAI工具，可以方便的使用预训练模型，如SqueezeNet，ResNet，InceptionV3 和 DenseNet
Step2，编写云函数在本地创建imageDemo项目
创建云函数入口文件 index.py
同时需要下载预训练模型：
- SqueezeNet，4.82 MB，预测时间最快，精准度一般
- ResNet50，98 MB，预测时间较快，精准度高（by Microsoft Research）
- InceptionV3，91.6 MB，预测慢，精度更高（by Google Brain team）
- DenseNet121，31.6 MB，预测较慢，精度最高（by Facebook AI Research）
Step3，下载python依赖工具
因为使用到imageai等工具，所以需要下载依赖，文件比较大，放到
链
接：https://pan.baidu.com/s/1MmhL3yHOXQiTN6x4Kg04iQ
提取码：7znq
Step4，部署AI应用 使用 sls --debug 部署后得到 API URL
Step5，本地调用API接口测试

腾讯广告算法大赛2019

总结Summary

- 记录作业内容
  - Thinking1：CTR数据中的类别数据处理，编码方式有哪些，区别是什么
  - Thinking2：对于时间类型数据，处理方法有哪些
  - Thinking3：你是如何理解特征组合的，请举例说明
  - Thinking4：DCN和xDeepFM都可以进行自动特征组合，有何区别
  - Thinking5：今天讲解的特征组合只是特征工程中的一部分，你理解的特征工程都包括哪些，不防做个思维导图
  - Action：Avazu CTR广告点击率预测：
    - 数据集：https://www.kaggle.com/c/avazu-ctr-prediction
    - 完成代码，提交submittion

- 说明采用的模型，LogLoss值
    - 鼓励交流，数据处理 or 模型使用（助教）

-

我觉得特征工程包括特征提取、特征组合以及特征选择；特征提取即通过PCA、ICA、LDA等方式将原始特征进行重新构建成新特征；特征组合即通过人工或自动的方式将不同特征间进行交叉组合，形成新的特征；特征选择则是在特征组合的基础上挑选与业务问题相关性强的特征组成训练集。同时，我认为特征工程的上述三个步骤实际是遵循PDCA原则的，即提出特征构想、形成新特征、通过简单模型验证其有效性，在此基础上优化特征并提出新的特征构想；通过不断地循环迭代不断接近最优的方案。