初步尝试

数据集

- 为了保持数据集的可管理性并保护与点击率、转化率（安装）相关的机密信息，对上述四个子集对应的记录进行**差分下采样。**
  - 差分下采样：差分隐私（一） Differential Privacy 简介 - 知乎 (zhihu.com)

    差分隐私系列之一：差分隐私的定义，直观理解与基本性质 - 知乎 (zhihu.com)
- 特征说明：每一行一共有80个特征（f_0 to f_79），两个标签（is_cilcked and is_installed）

  除去RowId和Labels，一共41个类别特征，38个数值特征

```
3. The data types of different columns are:
   a. RowId(f_0)
   b. Date(f_1)
   c. Categorical features(f_2 to f_32)
   d. Binary features(f_33 to f_41)
   e. Numerical features(f_42 to f_79)
   f. Labels(is_clicked, is_installed)
```

- 提交：RowId Labels（三列）
- 数据规模：20个训练数据集，csv文件，每个文件行数略有差别，列数均为82

  总共训练数据为(3485852,82)，平均每个文件(116195,82)

具体信息

- 训练数据由过去2周的二次抽样印象/点击/安装组成，旨在预测第15天的安装概率
- 评价指标：归一化交叉熵

$$Normalised\ Entropy = \frac{-\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1+y_i}{2}log(p_i)+\frac{1-y_i}{2}log(1-p_i)\right)}{-(p\times log(p)+(1-p)\times log(1-p))}$$

模型选择

deepFM

深入浅出DeepFM - 知乎 (zhihu.com)

(147条消息) 深度神经网络（DNN）模型*深度神经网络模型*ml_hhy的博客-CSDN博客

(147条消息) AI上推荐 之 FNN、DeepFM与NFM(FM在深度学习中的身影重现)_翻滚的小@强的博客-CSDN博客

embedding层的理解：(147条消息) 深度学习中Embedding层有什么用？_赵大寶Note的博客-CSDN博客

代码参考：

DeepFM全方面解析（附pytorch源码） - 知乎 (zhihu.com)

清晰易懂，基于pytorch的DeepFM的完整实验代码 - 知乎 (zhihu.com)

chenxijun1029/DeepFM_with_PyTorch: A PyTorch implementation of DeepFM for CTR prediction problem. (github.com)

服务器

ssh wyzhang@222.195.93.60 -p 1640

可能的一些问题

将所有数据整合在一起（包括测试集）处理缺失值并归一化，然后将所有分类特征统计value_counts总个数并重新编码，这种处理方式是否合适

click和install的结果如何得到，预测install的时候是不是要用到click，那么预测click的时候呢

模型调参

需要调整的参数

早停策略等

```
def __init__(self, feature_sizes, embedding_size=16,nume_fea_size=38,
             hidden_dims=[64, 64, 64], num_classes=1, dropout=[0.3, 0.3, 0.3],
             use_cuda=True,cuda_name="cuda:0"):
```

```
def fit(self, loader_train, loader_val, optimizer, epochs=100, verbose=True, print_every=100, wait=8,lrd=True):
    """
```

具体调参信息

1.

```
# training
model = DeepFM(feature_sizes, embedding_size=8,nume_fea_size=38,
               hidden_dims=[64,64,64],num_classes=1,dropout=[0.3,0.3,0.3],
               use_cuda=True,cuda_name='cuda:4')
# weight_decay 权重衰减系数 L2正则化项，防止过拟合，也可能导致模型训练效果下降
optimizer = optim.Adam(model.parameters(), lr=1e-2, weight_decay=0.0)
schedule = ReduceLROnPlateau(optimizer,'min',factor=0.2,patience=3,min_lr=1e-6,verbose=True)
print("Start Training...")
# wait用于早停策略
model.fit(loader_train, loader_val, optimizer,schedule, epochs=100,
          verbose=True,print_every=500,wait=12,lrd=True,figure_num=1)
# model.Get_result(loader_test,model)
```

epoch8，最终min_val_loss = 0.5195393638244985,Accuracy = 0.870927095413208

提交后：8.489126

改为test_size = 0.2，patience=4之后得到新的一轮结果并且提交测试

Epoch = 9,最终min_val_loss = 0.314258,Accuracy = 0.8705

score=

2.embedding_size = 4*

epoch6，最终min_val_loss = 0.3041278957751704,Accuracy = 0.8695070743560791

Epoch = 10,最终min_val_loss = 0.305031,Accuracy = 0.8698

3.embedding_size = 12

epoch8，最终min_val_loss = 0.5306406267282722,Accuracy = 0.8702844977378845

Epoch = 9,最终min_val_loss = 0.532616,Accuracy = 0.8701

4.embedding_size = 2

epoch5，最终min_val_loss = 0.304637,Accuracy = 0.8701

Epoch = 6,最终min_val_loss = 0.306314,Accuracy = 0.8689

5.embedding_size = 4,hidden_dims=[32,32,32]

epoch5，最终min_val_loss = 0.305936,Accuracy = 0.8695

提交后：12.476142

Epoch = 8,最终min_val_loss = 0.305789,Accuracy = 0.8695

6.hidden_dims=[16,16,16]

epoch6，最终min_val_loss = 0.306576,Accuracy = 0.8691

Epoch = 11,最终min_val_loss = 0.306971,Accuracy = 0.8691

7.hidden_dims=[128,128,128]*

最终min_val_loss = 0.303617,Accuracy = 0.8701

Epoch = 8,最终min_val_loss = 0.304270,Accuracy = 0.8701

8.embedding_size = 4,hidden_dims=[128,128,128],dropout = [0.1,0.1,0.1]

epoch6，最终min_val_loss = 0.299513,Accuracy = 0.8720

Epoch = 9,最终min_val_loss = 0.303845,Accuracy = 0.8705

9.dropout = [0.2,0.2,0.2]

epoch6，最终min_val_loss = 0.311635,Accuracy = 0.8702

Epoch = 9,最终min_val_loss = 0.303161,Accuracy = 0.8708

论坛信息

- 认为语义不足：Data --- 数据 (google.com)
- 不确定评估指标：Clarifications on the evaluation metric --- 关于评估指标的说明 (google.com)

  在我们的评估代码中，我们假设标签介于 0 和 1 之间。

  由于不同广告的安装概率差异很大，因此我们需要一种机制来规范基本点击率。分母中的 p 源自广告客户在过去 30 天内观察到的基本点击率。

  处理数值稳定性：我们使用一个值 ϵ，这样如果预测分数为零，我们使用 \epsilon 而不是 0，如果它是 1，那么我们使用 1 - ϵ而不是 1。

  我们只考虑评估is_installed，最终评估不考虑is_clicked指标。
- 提交文件的限制大小是25MB
- 根据前21天的信息预测第22天的安装概率？：Challenge Questions | Dataset & Evaluation --- 挑战问题 |数据集和评估 (google.com)
- 训练集和测试集采样：Test set distribution --- 测试集分布 (google.com)

  训练集和测试集都使用相同的采样算法进行采样。训练集基于连续 21 天进行采样，测试从第 22 天开始采样。
- 指标是基于完整集计算的
- 未完待续...