挥霍的人生


TAAC
腾讯广告算法大赛
Tencent Advertising Algorithm Competition

Contents

目录

团队介绍

团队介绍

周青松
哈尔滨工业大学
Kaggle Master

陈婷
深圳大学

罗宾理
中南大学
Kaggle Master

02

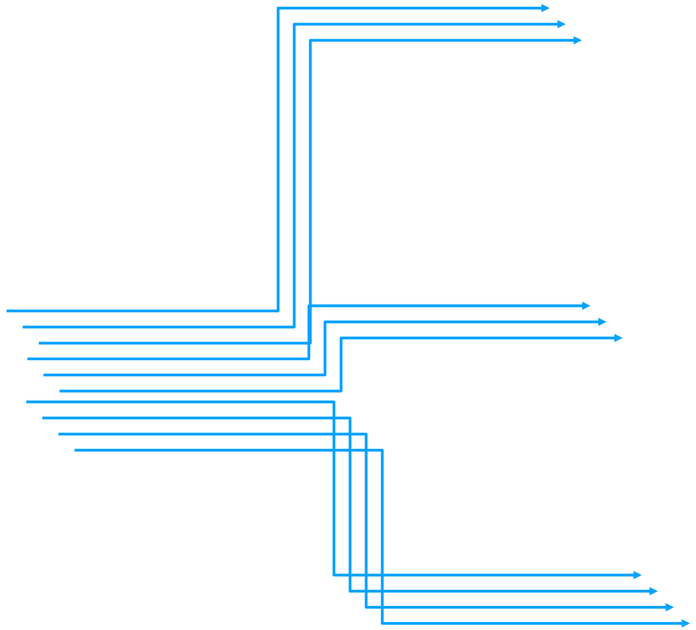
赛题理解

赛题理解

“参赛选手基于用户在广告系统中的交互行为作为输入来预测用户的人口统计学属性。”

“广度·人群智能定位
深度·人群属性校正”

Multi-Task
多任务学习



Node Classification
节点分类任务

Text Classification
短文本分类

CTR MultiClass
点击率预估多分类

特征工程

特征工程

Stats Feature

- 计数特征
- 对Count编码后的统计特征

Text Feature

- TF-IDF/CountVec 基学习器特征
- 词向量特征(Word2Vec / Fasttext / Glove)

Graph Feature

- ID间二分图
- 以ID为GroupBY下聚合的Sequence图
- ***DeepWalk / ProNE Embedding***
- ***Deep Graph InfoMAX (Attribute为每个ID的统计特征)***

模型介绍

文本分类模型

Model 1

Mix UP Embedding
(W2V Glove)
|
Embedding Layer
|
LSTM Layer
|
MaxPool Layer
|
Dense Layer

Model 2

Mix UP Embedding
(W2V Glove FastTEXT ProNE)
|
Embedding Layer
|
LSTM Layer + CNN Layer
|
Attention Layer
|
Dense Layer

Model 3

Mix UP Embedding
(W2V Glove FastTEXT ProNE)
|
Embedding Layer
|
Self Attention Layer
|
Attention + AvgPool Layer
|
Dense Layer

节点分类模型

Model 1

DGI Embedding + Stats Feature
|
LightGBM

Model 2

ProNE Embedding + Text Feature
|
CatBoost(Text Mode)

Model 3

Mix UP Embedding
(ProNE + DeepWalk + DGI)
|
2 layer GCN

05 总结与思考

分数相关

MixUP Embedding	↑0.002
不同模型	↑0.006
图特征	↑0.002

总结与思考

- 我们团队在此次比赛有些欠缺了尝试，一来是硬件配置的限制，二来是数据量的限制，在实验中经常爆内存，导致有很多想法没有实现出来。

- 此次数据量巨大，经分析发现数据是从9.1号开始持续三个月的，这里面包含了假期数据，但是没有合理的运用上这部分分析结果

THANKS