大作业汇报

刘寰硕, 2022/12/29

目录

摘要

1. 复现原论文以及示例pdf
    1. 给出选择不同λ以及不同数据集处理下的结果
    2. 最终复现的AUC和系数与示例pdf一致
2. 分析该数据集与推荐系统CTR任务的相似性
    1. 使用推荐系统中两个著名的baseline: DeepFM和xDeepFM给出更高精度的预测

复现/改进《最终研究成果2022》中的数据分析结果

数据预处理

遵循原论文以及示例pdf(TA在群里给出的报告)，我做出了如下预处理:

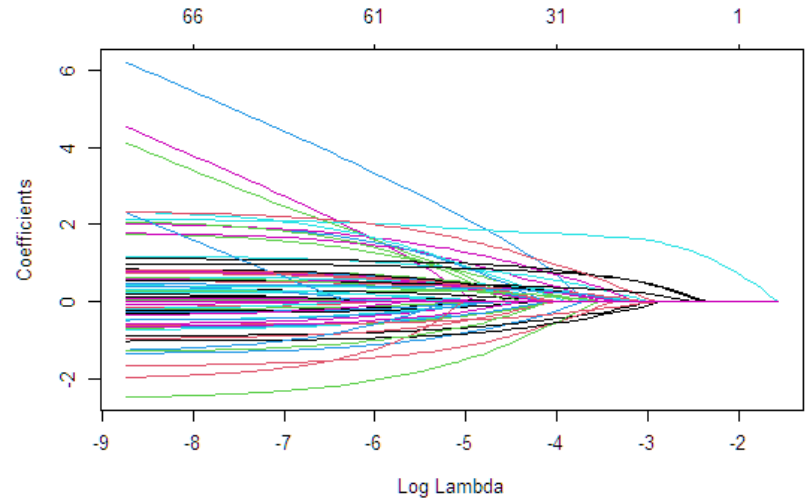
1. 保留group1和group2的样本，并将label分别记为0和1
2. 去除了肥胖/高脂两列特征
3. 去除了有缺失值的样本
4. (optional)将age和HR离散化 **(Note:对于区间的划分，示例pdf与原论文有差异，此处选择示例pdf的做法)**

数据集划分

1. 原论文:用$3:1$的比例划分训练集和测试集(不使用验证集)
2. 示例pdf采用的划分方式是:使用全部数据找出λ并筛选特征，使用k折交叉验证的方式计算AUC等指标
    1. 示例pdf中并没有具体阐明筛选特征时数据集的划分方式。因此此结论是与多组同学交流后得到的观点
3. 在本报告中，我会尝试使用示例pdf的方式得到了相似的结果，但为了避免模型过拟合，我在最终预测以及改进模型的时候时候采用原论文的划分方式。

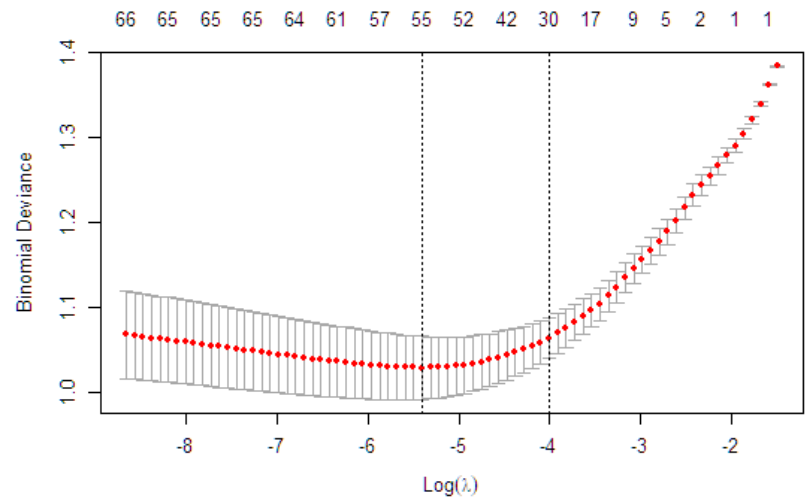
使用Lasso回归筛选变量

发现结果与原论文保持一致



使用Lasso回归筛选变量

发现结果与原论文保持一致



使用Lasso回归筛选变量

使用示例pdf中的$\lambda(\lambda = 0.022)$筛选变量,系数与示例pdf保持一致

features	coef
sex	.
age	0.002344185
LEF	0.074513999
HF_Af	-0.079014722
st	.
MI	.
pVTE	-0.382960723
SCI	.
AS	.
AID	.
BT	.
CVP	.
Chemo	.

使用Lasso回归筛选变量

对于预测，我们尝试了不同的setting。并得到了以下观察:

1. 当使用k折交叉预测时，模型精度接近原论文
    1. AUC: 0.807
2. 当使用所有数据进行训练和预测时，模型精度接近示例pdf
    1. AUC: 0.821
3. 当划分训练集、测试集时由于测试集样本量较小，模型的AUC反而会比其他两种Setting更大
    1. AUC: 0.83左右

应用其他方法提高预测精度

我们分别使用DeepFM和xDeepFM进行预测。这两个模型是推荐系统领域比较流行的Baseline。

动机

对于本任务，我们可以把任务抽象成:

1. 使用高维稀疏分类数据对样本进行二分类(是否是DVT+APE)
2. 将DVT+APE的样本记为正例，其余样本记为负例。

这一点与推荐系统领域中的CTR预测任务十分接近

CTR预测任务

CTR预测任务的目标是计算出某类用户在给定环境下购买给定商品的概率有多高，例如要给特定用户推荐某个或者某些电影（电影是一个产品），这个用户看这个电影的概率有多高。它具有如下几个特征:

1. 大量离散特征。例如性别(男,女)以及学历(无学历，小学，中学，高中，中专，大专，本科，硕士，博士)。

2. 大量高维度稀疏特征

    1. 高维度指的是样本量和特征维度大
    2. 稀疏指的是在数据中可能只有少数数据是1。

3. 特征工程(特征组合)十分重要

    1. 在电商网站系统中，需要预测一个用户是否购买一个东西，会是什么样的特征组合。性别是一个特征，时间是一个特征，所谓的特征组合就是假设性别特征是女，时间维度双11。具体实践中，我们会发现这两个特征的组合会带来十分强烈的购买倾向。例如女性在双十一购买商品的概率非常高，这就是组合特征。组合特征对CTR挖掘非常关键。

与我们的任务进行对应

接下来我会将上述特征与我们的任务一一对应起来。

1. 对于特征1,我们的数据集中只有两个特征不是分类特征(Age和HR，但在后续处理中也会将它们离散化)

2. 对于特征2，我们数据的稀疏度(Sparsity)为99%

    1. 这种稀疏程度甚至比多数推荐系统公开数据集还要高

3. 对于特征3,通常医生在预测一个病人是否患病的时候需要考虑多种因素，当多种症状同时指向一种病症时才可以做出把握性比较大的判断

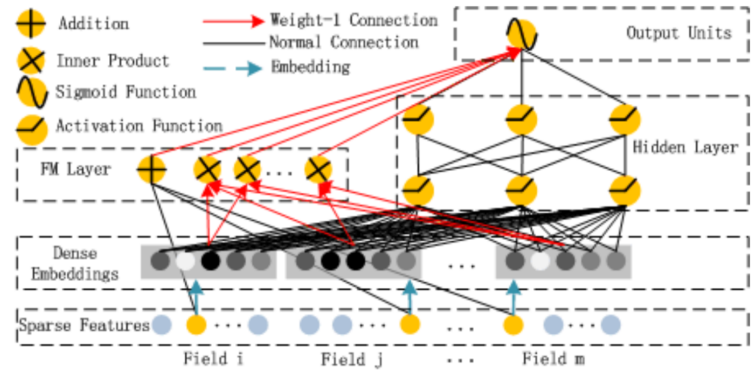
4. 当然，相比于推荐系统动辄上亿的数据量，我们的数据相对来说比较少，可能会出现过拟合风险，这一点需要用实验进一步验证

DeepFM

1. 由于xDeepFM与DeepFM十分接近，因此我只选取了后者进行介绍

2. DeepFM使用了因子分解机FM加强浅层网络部分特征组合的能力

    1. 左侧的FM部分与右边的深度神经网络共享相同的Embedding层
    2. 左侧的FM对不同特征域的Embedding进行了两两交叉(将Embedding向量当作原FM中的特征印象里)
    3. 最后将FM的输出域Deep部分的输出一同输入最后的输出层，参与最后的目标拟合



DeepFM架构图

最终结果

可以知道，深度学习带来的提升比较明显。这主要是得益于选用的模型对特征进行显式的特征交互，并且使用了MLP增强了模型拟合能力。

Model	Lasso+Logistic	DeepFM	xDeepFM
AUC	0.8326	0.8472	**0.8523**

表 5: 深度学习预测结果

预测结果