知识图谱工程-项目作业 2

Nov 6, 2018

目标：基于知识图谱，搭建风控模型来预测一个进件的逾期与否。

本项目提供了以下的数据文件：

- person.txt 包含了每个申请人的属性 （跟作业 1 一样）
- phone.txt 每一行记录标记给定的电话号码是否在黑名单里
- phone2phone.txt 详细的通话记录，包含通话时长，通话时间等
- apply_train.txt 包含了进件信息，另外最后一个字段是进件的状态，这个数据用于训练模型
- apply_test.txt 包含了进件信息，但这里并没有包含进件的状态，这个数据用于测试模型。

对以下几个属性做简单说明：

- amount: 申请人所申请的贷款额度
- term： 还贷期限（比如 20 个月）
- status: 进件状态
- flag: 黑/白名单

如果跟作业 1 相比较的话，数据方面有几点不一样：

- 新的数据文件 phone.txt，里面包含每个电话的标签
- apply.txt 分成了两部分，分别是 apply_train.txt 和 apply_test.txt，它俩的区别是 apply_test.txt 不包含进件的状态
- person.txt, 这个文件也多了一个属性，就是对于一个人的标签（白/黑）。黑代表被标记为黑名单。

任务 1： 通过给定的数据来搭建知识图谱，这个过程类似于作业 1。请把知识图谱的设计图存成 design.png。

本次作业的核心是搭建风控模型，来预测某一个进件的状态是否是逾期。对于一个进件，它有多种状态，OVERDUE(逾期)，IN_PROGRESS(正在审核中)，RETURNING(偿还中)，REPAID(已偿还)。为了方便起见，我们可以把 IN_PROGRESS, RETURNING, REPAD 均看做是非逾期的进件或者正常进件（NORMAL）。也就是说我们需要搭建一个二分类器来预测一个进件是否会逾期或者不会逾期。

所以，对用于训练的进件，我们需要提取一些特征，那这些特征有可能是进件本身的，也有可能是从关系网络里提取出来的。对于一个进件，提取完特征之后，我们即可以得到如下的训练数据，例如：

进件 1，特征 1，特征 2，特征 3，…., 特征 30，OVERDUE
进件 2，特征 1，特征 2，特征 3，…, 特征 30，IN_PROGRESS
进件 3，……..

那基于这样的训练数据，我们可以训练一个机器学习模型来预测一个进件的逾期与否。

任务 2： 设计一套有效的特征。那这种特征可以分成两大类：1. 基于规则的特征 2. 直接提取出来的数值类型特征（具体详见第五章）。创建 feature.txt 文件，把每一个特征描述加进去。所以格式为每一行一个特征描述，如(给出了三个特征)：

进件的申请人之前有没有逾期过
二度关系中触碰黑名单的个数

二度关系中触碰黑名单的电话个数

…..

任务 3： 对于任务 2 中所选择的特征，把特征提取过程写成 cypher 语句并写到 mysql 数据中（具体详见第六章）。把整个数据库导出为 hw2.sql（跟第一次作业一样）

任务 4： 根据 mysql 里面的 cypher 语句，在 java 工程里编写调用 cypher 语句的过程，此过程请写成微服务（具体详见任务第六章）。并写一个 java 的脚本来整合所有的特征（最后就可以得到类似于前面所提到的训练数据格式）。

任务 5：利用逻辑回归，GBDT, 神经网络在训练数据上搭建风控模型, 这是二分类问题（逾期/非逾期）。评估标准请用 AUC 指标。在训练模型时需要利用交叉验证技术来选择最优的超参数（请使用 5-fold cross validation）。对于每一组超参数，请尝试 10 组不同的 5-fold cross validation。每次 5-fold 交叉验证我们会得到一个平均准确率，循环 10 次之后，我们就可以得到这 10 次的平均值还有标准差。请生成一个表格来填写模型的准确率（AUC 值），分别为平均值和标准差。表格的格式为：

模型	训练数据上的 AUC（平均准确率）
逻辑回归	0.8（平均值） ± 0.04（标准差） --- 举个例子
GBDT	…..
神经网络（一层 hidden layer）	…..

把这表格存成一个图片 train_auc.png

任务 6： 通过交叉验证之后，请选择一个最好的模型，然后对于 apply_test.txt 中的进件做预测(逾期/非逾期)，并把结果写到一个新的文件 apply_test_pred.txt, 每一行为对应的预测值。写入的时候，如果预测为逾期的，那就写字符串“OVERDUE”, 如果预测为非逾期的话就写字符串“NORMAL”。 apply_test_pred.txt 的格式为：

label
OVERDUE
NORMAL
NORMAL
NORMAL
OVERDUE
NORMAL
……

测试进件的真实的状态在助教的手里，所以提交完 apply_test_pred.txt 之后，我们会通过一个简单的脚本来去计算模型在测试数据上的准确率，并根据这个准确率来给任务 6 打分。所以，提交时务必要保持顺序的一致（跟 apply_test.txt 中的进件顺序保持一致）。提交前，再次确认，apply_test.txt 文件里的行数是否等于 apply_test_pred.txt 中的行数。

作业的提交：把以下的文件打包成 [账号_hw2.zip]，发给助教。

- design.png
- feature.txt

- hw2.sql
- train_auc.png
- apply_test_pred.txt