R&S | 手把手搞推荐[6]：回顾整体建模过程

原创 机智的叉烧 CS的陋室 2019-07-14

Heart Like California

Before You Exit - Heart Like California

往期回顾：

- R&S | 手把手搞推荐[4]：打分预估模型
- NLP.TM | 命名实体识别基线 BiLSTM+CRF（下）
- 我的半年总结
- 算法与数据分析秋招经验【送内推码】
- NLP.TM | tensorflow做基础的文本分类

连同之前的"评价指标设计"，转眼手把手搞推荐已经更新5期，前面一段时间内，详细地阐述一个简单的推荐系统模型建立的过程，从问题分析、数据探索、数据处理、特征工程乃至建模角度，说了是手把手，就给大家揉碎了讲，谈自己的经验和看法，同时坚持给大家看到代码，如果大家能认真看完，认真理解，相信就会有很大收获，所谓"为故而知新，可以为师矣"，本文带着大家重新看看，无论是从推荐系统本身而言，还是一个算法项目而言，都需要经历什么思考以及什么具体工作。

对了，有关本模块的代码，都在这个码云链接下：

https://gitee.com/chashaozgr/noteLibrary/tree/master/rs/src/LR

欢迎大家进去看看，有问题欢迎提issue，另外这个项目下有些有关我的公众号的代码也在里面，大家也可以进去逛逛。

0 入门小结

R&S | 手把手搞推荐[0]：我的推荐入门小结

第一篇文章详细讲了我的入门过程，写了一些我入门的经验，相信这些经验能够给到大家不小的帮助。

- 简单的路线推荐：机器学习基础、深度学习基础、召回方法、排序与CTR预估、大型推荐系统架构、论文深入和案例练习，期间要加入embedding、冷启动等问题的学习
- 基础的学习一般都很容易找到资料，但是要进阶的话还是会遇到瓶颈，建议多看论文以及各个大厂的报告，例如datafuntalk以及一些大会，会有公司给出自己公司的一些方案，大家都多看看理解一下，有条件的可以自己尝试重现，论文和报告才是前沿。
- 多动手，我说了很多次，这里再次强调。

1 数据探索

R&S | 手把手搞推荐[1]：数据探索

数据探索是一个很容易忽视的过程，但却是不能忽视的，这是一个结合实际问题探索数据中的规律的过程，我在这篇文章中以movielens为例讲解了数据探索的思维过程。

- 目标确定：确定你要做的实际问题是什么。
- 需要探索的内容：有什么数据、格式如何、质量如何，然后分析数据的分布，有无周期性等性质，在做分类时还要分析是否有数据不平衡的情况，
- 进行数据整理，尝试把数据转化为适合存取等方式的结构，方便后面计算和分析。

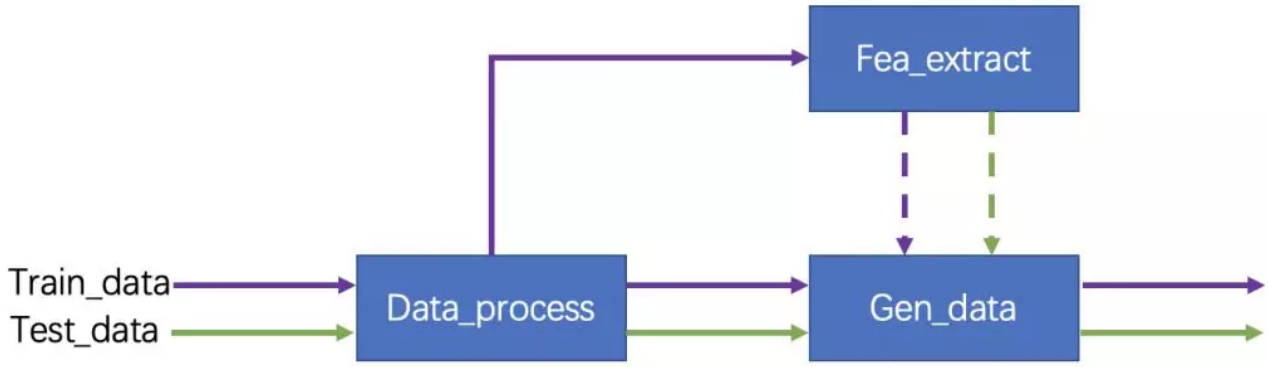
在文章总有我提到的详细代码，可以点进去看看的。

2 特征工程

R&S | 手把手搞推荐[2]：特征工程指南

俗话说"特征选不好，调参调到老"，特征工程可以说是一个体力活，但是却需要对问题有充分的了解，只有理解特征，才能在选好特征，有用的特征有利于最终的预测结果，拿着无关的特征在改模型肯定是没有意义的。

这里详细说到了一个特征工程模块设计的思路，包含数据处理、特征抽取、数据生成三个部分，如下图所示。



Dataprocess是一套基本的数据操作，把一些内部的特征提取出来，以及一些异常数据剔除或者转化，做的工作主要就是一些特征上的基本操作，这些操作是**是否把训练集测试集分开都能做**的工作。Feaextract是特征提取工作，这套工作只需要训练集数据完成即可，目标是构建一些特征提取的算法。Gendata是数据生成工作，根据Feaextract中对每个特征处理建立的函数，生成转化后的数据集，这块工作当然训练接集和测试集都需要进行。

文章后半部分，开始给大家讲这三步如何实现，具体代码可以点击原文去看。

3 数据集存取

R&S | 手把手搞推荐[3]：数据集存取思路

在特征工程里，其实提到了特征工程后的数据存放问题，数据库等是一个比较可行的方式，但是由于数据基本只用作训练和测试，没有具体很多功能，所以越是简单的方式越好，只需要方便整体存入和整体取出即可，此处考虑使用稀疏矩阵的方式存储，格式为npz方式，方便存取。

文章后半部分同样代码实现，具体可以点进去原文查看。

4 打分预估模型

R&S | 手把手搞推荐[4]：打分预估模型

有了用于训练的数据集，就可以开始进行模型训练了，但是，选什么模型，怎么选，肯定不是每个模型都去试试然后得到结果的，如果能通过问题本身的性质和各个模型之间的关系进行分析和诊断，会节省大量的时间，这个无论在竞赛中，还是在实际问题分析下，都是非常重要的。在本次讨论中，我选择的是逻辑斯蒂回归模型，主要从下面几个方向考虑：

- 第一个版本的模型要求尽可能简单，快速完成功能

- 测试或预测阶段效率尽可能高(毕竟涉及上线，响应时间非常重要)，KNN之类的肯定要抛弃
- 准确性达到基本标注
- 当前数据集内特征大都是离散型特征(已经全部onehot化了，就全都是离散特征了)

评价指标设计

评价指标设计

这篇虽然没有带上"手把手搞推荐"的抬头，但是其实是沿着这个线路去讲的，数据上和流程上都是如此。

首先谈到了我对指标的分类，个人主要把它分为性能指标和业务指标，两者有联系但又有区别，另一方面也就"对不对"和"差多远"两种情况来讨论具体问题下应该怎么去设计指标，有的时候我们需要分出对错，例如疾病诊断，要是有病得治，没病就不治，但是也有一些时候是看距离正确的多远，例如股票价格预测，基本不会直接命中，而是得到一个比较接近的结果，通过分析与实际的差距来进行检测。

然后，我在上面讨论的基础上，给出一种我自己去构造指标、分析指标、改进指标的思路历程，供大家参考，给定的指标只能解决给定的问题，大部分问题确实能够抽象从而用到这些指标，但是对于具体的问题，我们应该去构造针对这个问题的指标。

后续计划

根据自己的学习计划，在"手把手搞推荐"这个系列下，可能会有下面的内容，但是由于自己后续就要入职工作等原因，本系列的更新速度不会之前那么快：

- 有关embedding与召回方面的策略
- badcase发现、诊断与处理
- 排序模型的应用与分析
- 模型的上线与部署

当然的，这不代表本身"R&S"系列下就没有别的东西了，我依旧会在推荐和搜索上和大家分享我最近所学和发现，如推荐搜索方面的策略、经典论文、优秀的分享活动笔记等，敬请期待！