机器学习工程师 -- 开题报告

项目背景

随着互联网的发展，人们正处于一个信息爆炸的时代。相比于过去的信息匮乏，面对现阶段海量的信息数据，对信息的筛选和过滤成为了衡量一个系统好坏的重要指标。一个具有良好用户体验的系统，会将海量信息进行筛选、过滤，将用户最关注最感兴趣的信息展现在用户面前。这大大增加了系统工作的效率，也节省了用户筛选信息的时间。

搜索引擎的出现在一定程度上解决了信息筛选问题，但还远远不够。搜索引擎需要用户主动提供关键词来对海量信息进行筛选。当用户无法准确描述自己的需求时，搜索引擎的筛选效果将大打折扣，而用户将自己的需求和意图转化成关键词的过程本身就是一个并不轻松的过程。

在此背景下，推荐系统出现了，推荐系统的任务就是解决上述的问题，联系用户和信息，一方面帮助用户发现对自己有价值的信息，另一方面让信息能够展现在对他感兴趣的人群中，从而实现信息提供商与用户的双赢。

问题描述

传统的推荐方法主要包括协同过滤、基于内容的推荐方法和混合推荐方法。其中，最经典的算法是协同过滤，如矩阵因子分解，其利用用户与项目之间的交互信息为用户产生推荐，协同过滤是目前应用最为广泛的推荐算法，但是同时也遭遇到了严重的数据稀疏（一个用户评分过的项目仅仅占总项目数量的极少部分）和冷启动（新的用户和新的项目往往没有评分数据）问题。

近年来，深度学习在图像处理、自然语言理解和语音识别等领域取得了突破性进展，已经成为人工智能的一个热潮，为推荐系统的研究带来了新的机遇。一方面，深度学习可通过学习一种深层次非线性网络结构，表征用户和项目相关的海量数据，具有强大的从样本中学习数据集本质特征的能力，能够获取用户和项目的深层次特征表示。另一方面，深度学习通过从多源异构数据中进行自动特征学习，从而将不同数据映射到一个相同的隐空间，能够获得数据的统一表征，在此基础上融合传统推荐方法进行推荐，能够有效利用多源异构数据，缓解传统推荐系统中的数据稀疏和冷启动问题。基于深度学习的推荐系统研究目前已经成为推荐系统领域的研究热点之一。

本项目使用文本卷积神经网络，并利用MovieLens数据集完成电影推荐的任务。

输入数据

本项目使用的是MovieLens 1M 数据集，这些数据集大约包含6040个用户在3900部电影上的1000209个匿名评级。

数据集分为三个文件：用户数据users.dat，电影数据movies.dat和评分数据ratings.dat。

解决办法

通过研究数据集中的字段类型，发现有一些是类别字段，通常的处理是将这些字段转成one hot编码，但是像UserID、MovieID 这样的字段就会变成非常的稀疏，输入的维度急剧膨胀，这是不愿意见到的，所以在预处理数据时将这些字段转成了数字。电影类型的处理要多一步，有时一个电影有多个电影类型。电影名的处理比较特殊，没有使用循环神经网络，而是用了文本卷积网络。

基准模型

项目使用Tensorflow构建模型，训练出用户特征和电影特征，在实现推荐功能时使用。得到这两个特征以后，就可以选择任意的方式来拟合评分了。

评估指标

本项目使用MSE作为评估指标。

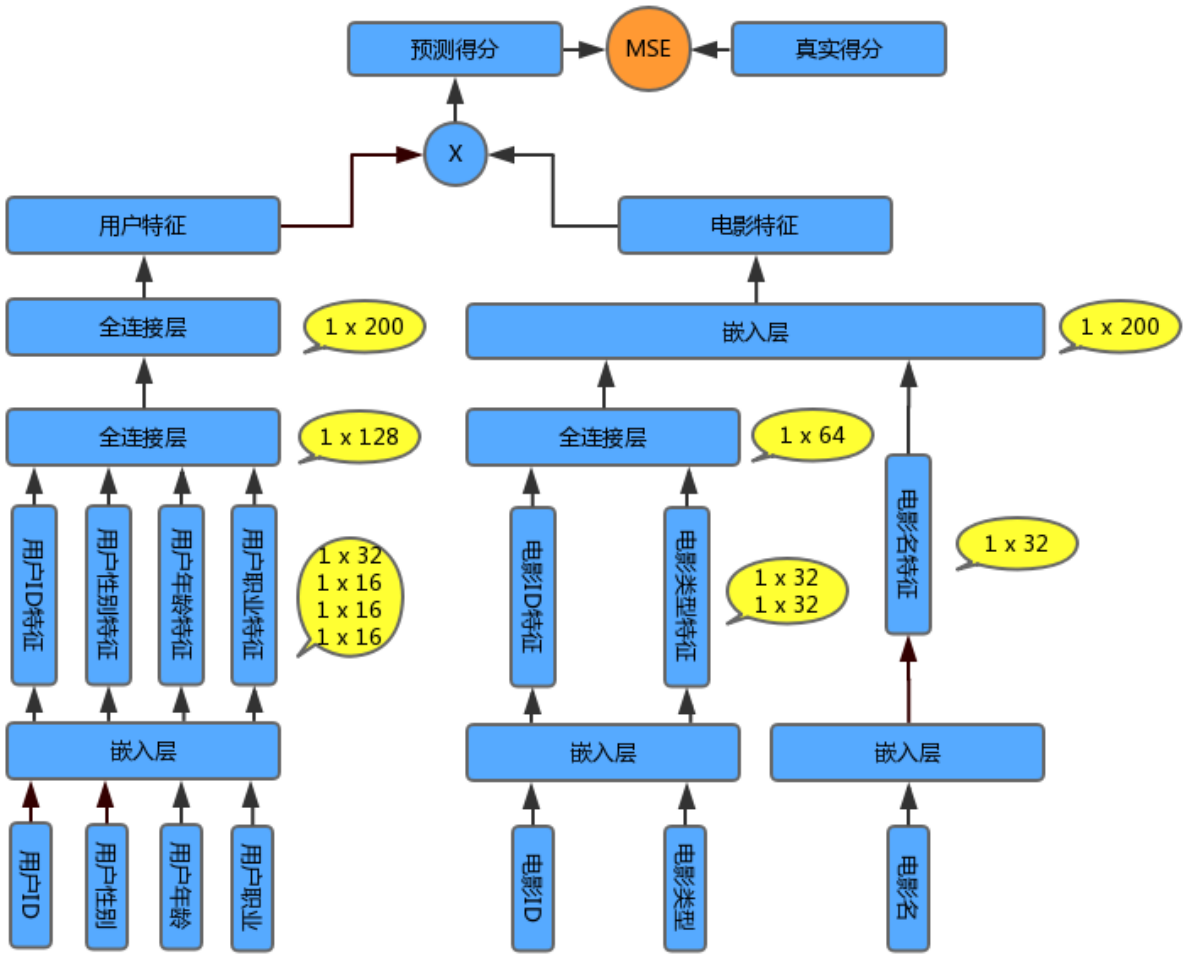
平均平方误差MSE（Mean Squared Error）又被称为l2范数损失（l2-norm loss），是反映估计量与被估计量之间差异程度的一种度量。

公式如下：

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}}\sum_{i=1}^{n_{sample}}(y_i - \hat{y}_i)^2$$

MSE可以评价数据的变化程度，MSE越小，说明模型的拟合实验数据能力强。

设计大纲



数据集经过处理后，用户特征网络第一层使用了嵌入层，维度是（N，32）和（N，16）。电影有多个电影类型，这样从嵌入矩阵索引出来是一个（N，32）的矩阵，因为有多个类型，需要将这个矩阵求和，变成（1，32）的向量。电影名的处理使用文本卷积网络得到（N，32）电影名特征。

从嵌入层索引出特征以后，将各特征传入全连接层，将输出再次传入全连接层，最终分别得到（1，200）的用户特征和电影特征两个特征向量。

实现的推荐功能如下：

- 指定用户和电影进行评分
- 推荐同类型的电影
- 推荐您喜欢的电影
- 看过这个电影的人还看了（喜欢）哪些电影