## 推荐系统第三次作业——隐语义模型

### 算法简介

#### **简介**

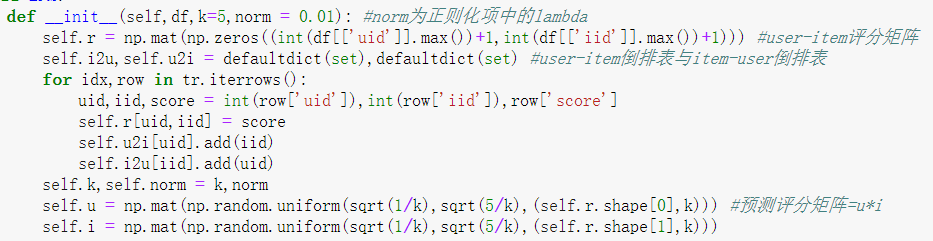
隐语义模型是推荐系统问题中的经典算法，对于user-item评分矩阵，假设每个用户和每个物品均可表示为一个k维向量，那么user-item评分矩阵可以被分解为用户特征矩阵和物品特征矩阵的乘积。定义具有L2正则化项的损失函数，用梯度下降或交替最小二乘法进行求解，即可求出用户特征矩阵和物品特征矩阵。

#### **实现过程描述**

首先套用lab1对数据的预处理过程，然后定义LFM类表示LFM模型。首先用DataFrame构造LFM对象，然后初始化user-item矩阵R与用户特征矩阵U和物品特征矩阵I，再建立user-item倒排表与item-user倒排表以加速训练过程。随后根据公式定义损失函数与两种不同的训练方法，再进行测试调整超参数，最后得到预测结果。

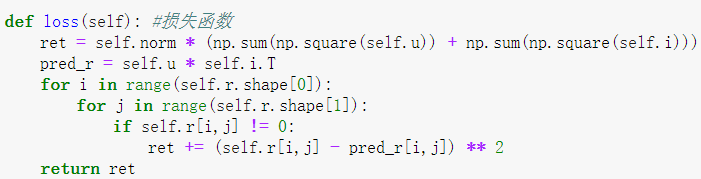
### 核心代码注解

#### **2.1 LFM类的构造函数**



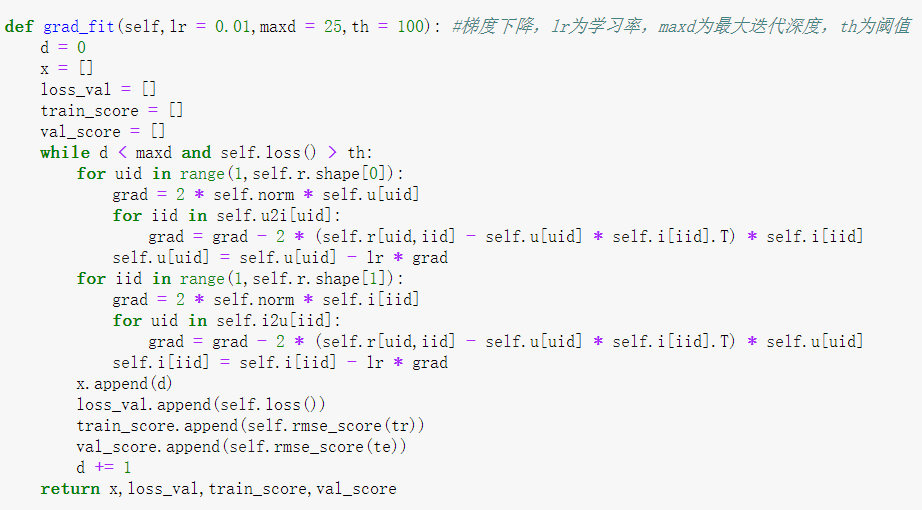
#### 首先初始化评分矩阵与两个倒排表，然后初始化k和正则化系数，最后随机初始化用户特征矩阵U与物品特征矩阵I。由于评分的范围为，要使得任意用户特征向量和物品特征向量的内积落在这个范围内，则必有和的元素范围为。

#### **2.2 损失函数**



首先根据公式计算两个正则化项的数值，然后找到所有训练集中的评分，计算分差的评方加和到返回值中。

#### **2.3 梯度下降**

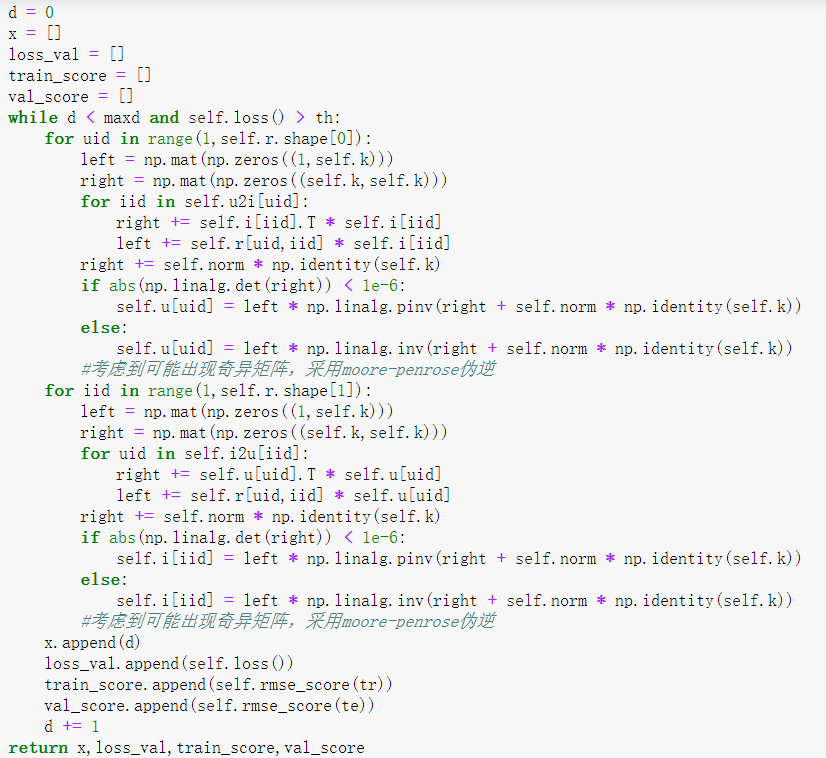


记录训练过程中的训练次数、损失函数值、训练及得分、测试集得分并返回，便于找到最佳的超参数。根据如下所示的公式（由于矩阵的含义不完全相同，可能和课件中的公式略有差别）计算梯度：



然后根据学习率、最大迭代次数、损失函数阈值进行梯度下降，得到训练结果。

#### **2.4 交替最小二乘**

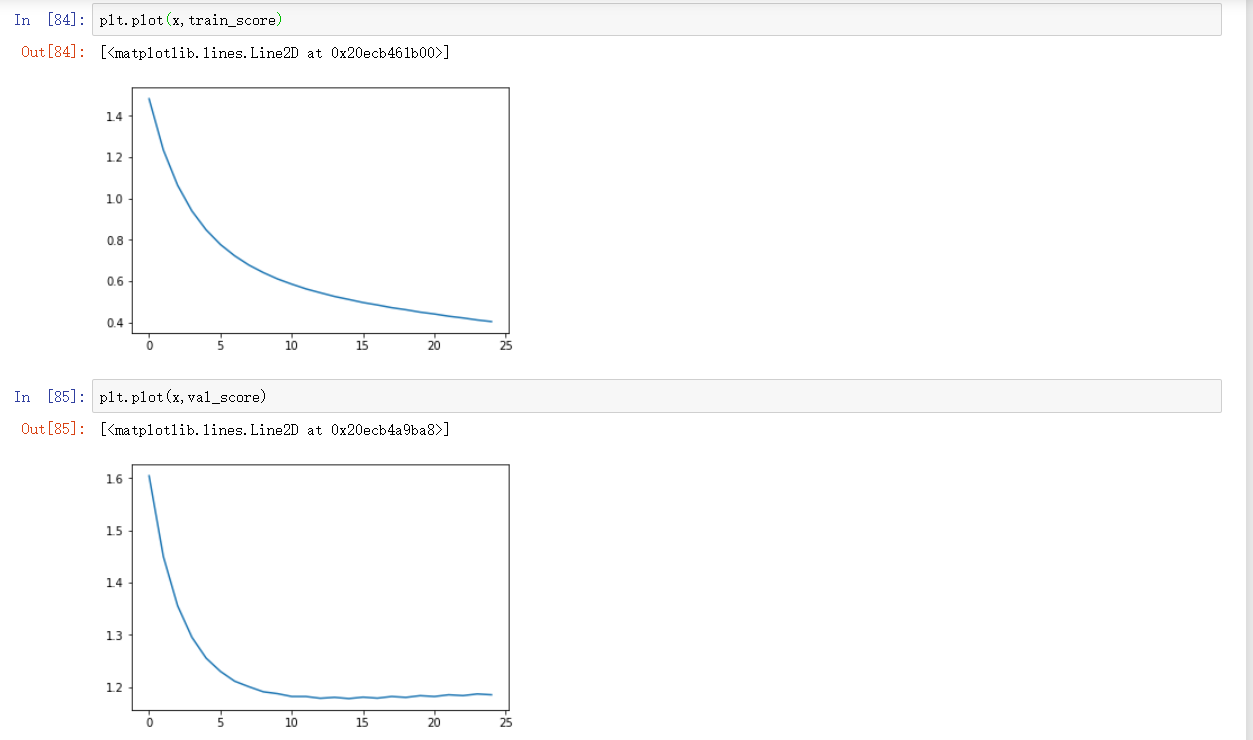
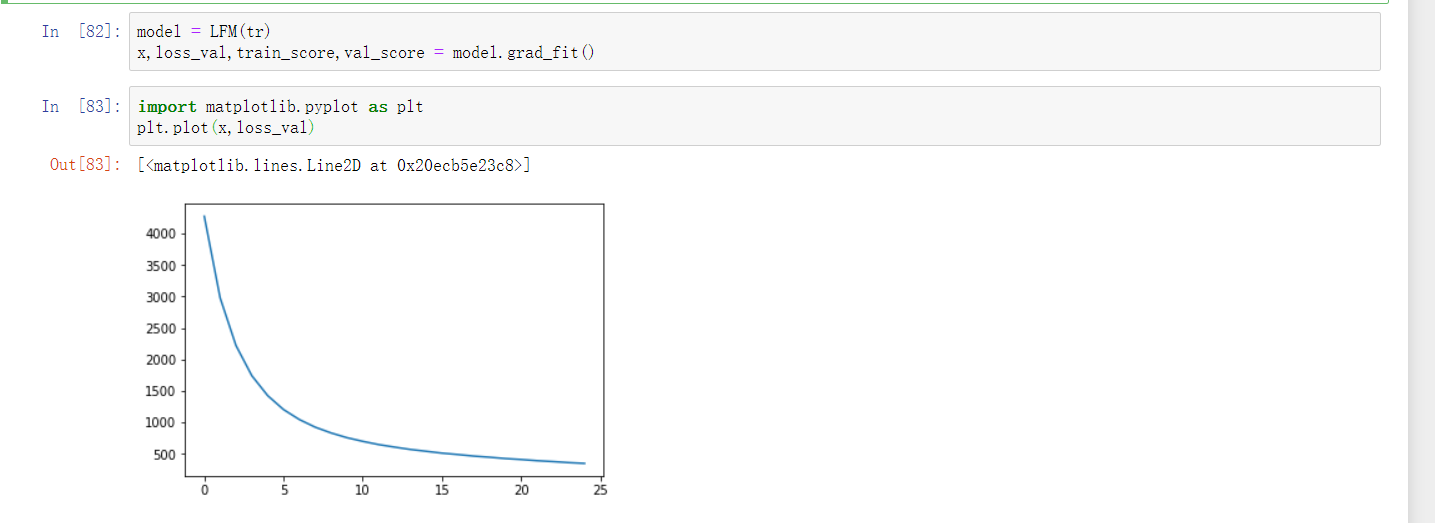


记录训练过程中的训练次数、损失函数值、训练及得分、测试集得分并返回，便于找到最佳的超参数。根据如下公式（由于矩阵的含义不完全相同，可能和课件中的公式略有差别）进行迭代：



考虑到奇异矩阵问题，在right矩阵为奇异时用Moore-Penrose伪逆代替普通的逆矩阵。然后根据最大迭代次数、损失函数阈值进行梯度下降，得到训练结果。

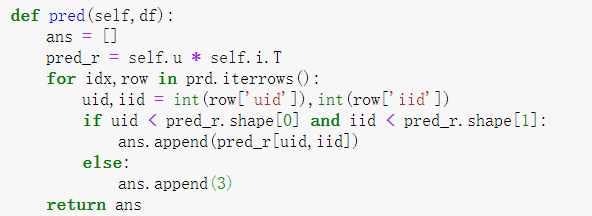
#### **2.5 超参数测试**



测试als\_fit和grad\_fit的效果，可以发现grad\_fit效果较好一些。用matplotlib画图，寻找合适的超参数，最终得到最合适的超参数如下所示：

k=5 正则项系数0.01 学习率0.01 最大深度15

#### **2.6 填写答案**



定义如下函数得到df中测试集的结果，最后将答案输出到csv文件中。



### 结果分析

模型最终在test集上得到的RMSE为1.1847390114315492，与lab1相比结果近似。

### 代码运行

#### **4.1 环境**

Anaconda 2.6.0

Python 3.7.1

#### **4.2 依赖**

numpy 1.16.2

pandas 0.23.4

matplotlib 3.0.2

#### **4.3 命令行命令**

命令行运行jupyter notebook 后打开本文件

### 提交文件

RS\_HW3.ipynb 本次作业

my\_answer.csv 我的预测答案

train.csv test.csv pred.csv 数据集