Data Mining, Spring 2018

Problem Set #3: PCA, Recommender System, and Association Analysis

(Due on June 24, Sunday)

**Submission Instructions**

These questions require thought but do not require long answers. Please be as concise as possible. We do not do reverse engineering, so please DO NOT provide MATLAB (or other programming language) codes WITHOUT **method description**. You should also declare in the assignment that **the MATLAB (or other programming language) code was written by you, not by others either partially or fully**.

You should submit your answers as a write-up in PDF format to [DataMining\_2018@126.com](mailto:DataMining_2018@126.com). The email title is formatted as “hwk3\_学号\_姓名”.

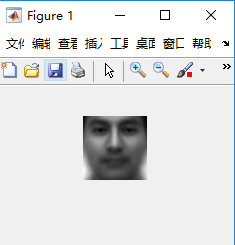
If you are in doubt, talk to me ([majh8@mail.sysu.edu.cn](mailto:majh8@mail.sysu.edu.cn)) or our teaching assistants to understand more.

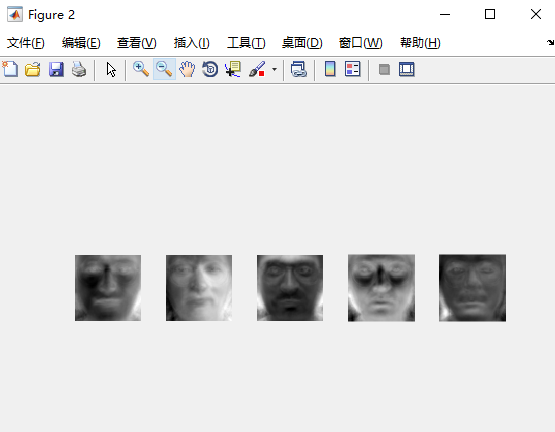
**Questions**

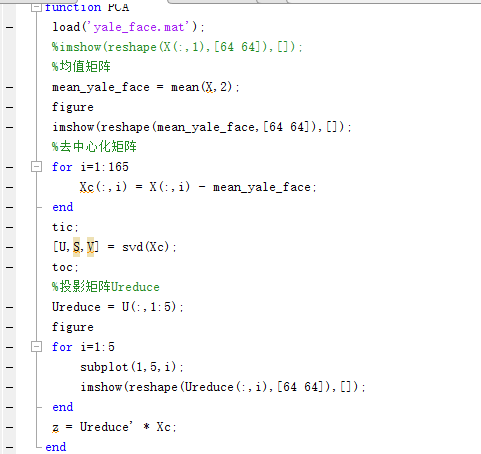
1. **主成分分析（Principal Component Analysis，PCA）**

请从课程网站或[此链接](https://pan.baidu.com/s/1djZy69OmLWqJDpStjGcQaA)下载Yale人脸数据集进行降维。通过MATLAB命令load('yale\_face.mat')读取数据，包含一个矩阵。（请注意，这里的矩阵X是课件第21页定义的数据矩阵X的转置。）此矩阵的每一列是由一张的灰度人脸图像所转成的维向量，即每一维列向量是一个训练样本，特征的维数是，样本的个数是。例如，可以使用imshow(reshape(X(:,1),[64 64]),[])命令显示第一张人脸图像（第一个训练样本）。

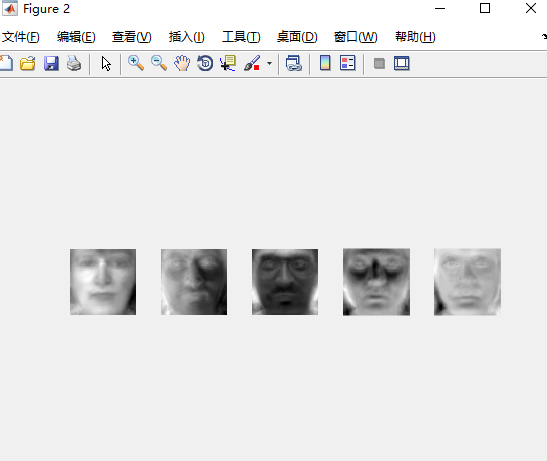
1. 试使用MATLAB中的svd函数实现PCA算法，即输入数据矩阵X和降维后的维数k，对每一个样本进行去中心化，然后对进行去中心化后的数据矩阵Xc用svd函数[U,S,V] = svd(Xc)，输出降维的投影矩阵Ureduce（即U的前k列），降维后的坐标表示Z=Ureduce’\*Xc，训练样本均值。并令k=5，显示样本均值的图像和Ureduce的五个列向量（即协方差矩阵的前五个特征向量）所对应的图像；







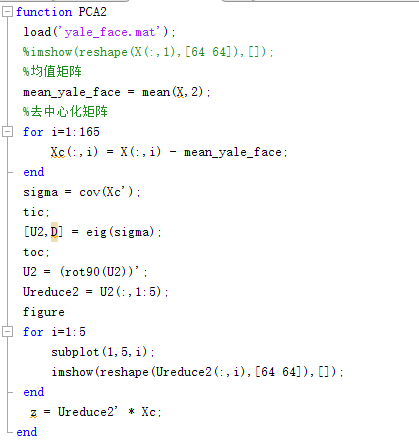
1. 试对协方差矩阵使用MATLAB中的eig函数计算特征值和特征向量，即[U,D]=eig(Xc\*Xc’/m)，显示前五个最大的特征向量所对应的图像，并比较对数据矩阵使用svd函数的所得出的特征向量的图像与运算时间；



svd: 

eig: 

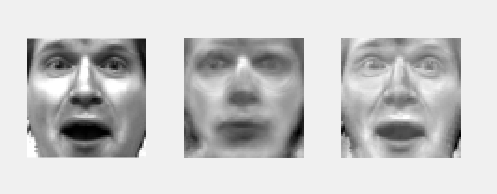
明显可以看出svd函数运行速度比eig快很多。

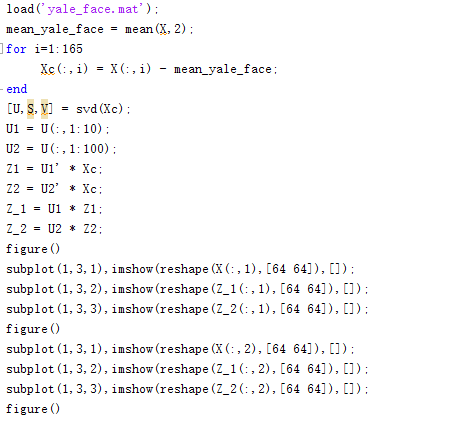


1. 试计算当降维后的维数分别是10和100时，保留的方差的比例（即或使用（1）中的S计算或使用（2）中的D计算），并分别利用10维和100维坐标恢复原高维空间中的人脸图像，对前三张人脸图像，对比原图和两张恢复的图像。

第一张图：（从左到右分别是原图，10维恢复图，100维恢复图，下同）







1. **推荐系统（Recommender System）**

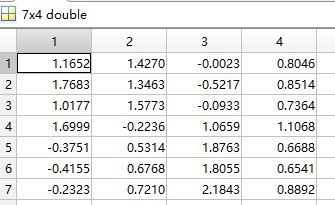
考虑以下的8个用户（A-H）对7部电影评级（1到5级）的一个效用矩阵：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E | F | G | H |
| HP1 | 4 | 4 |  |  | 1 | 1 | 5 |  |
| HP2 | 5 | 5 |  | 1 |  |  |  |  |
| HP3 |  | 4 | 1 |  |  | 1 | 5 | 4 |
| TW | 5 |  | 2 | 5 |  | 1 | 2 |  |
| SW1 | 1 |  | 5 | 4 | 5 |  |  | 1 |
| SW2 | 1 |  | 5 |  |  | 4 |  |  |
| SW3 |  | 1 |  | 5 |  | 5 | 1 |  |

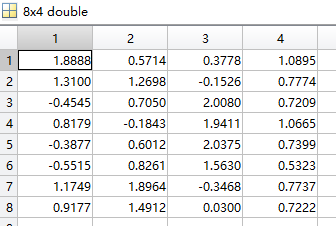
电影的名字HP1、HP2、HP3分别代表《哈利波特》（Harry Potter）I、II、III，TW代表《暮光之城》（Twilight），SW1、SW2和SW3分别代表《星球大战》（Star Wars）I、II、III。

1. 试实现协同过滤算法（Collaborative ﬁltering algorithm，课件第19页，不需要进行去均值操作），令正则化参数，特征向量维数，学习率，分别计算描述电影特征的矩阵和预测用户评级的模型参数矩阵（定义见课件第23页，所得结果保留小数点后4位），并计算预测电影评级的效用矩阵即（保留小数点后1位）；

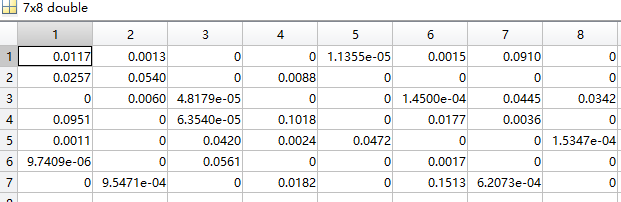
X：

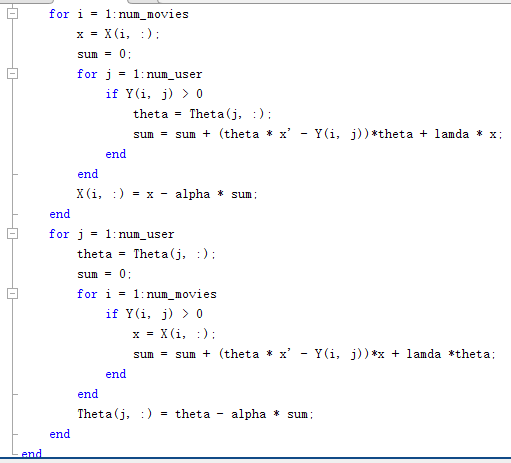


Θ：



Xθ：

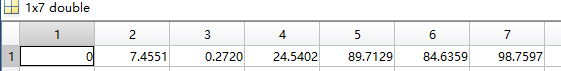




1. 试计算（1）中预测的电影评级与真实评级的平方误差，即，其中的定义见课件第7页；讨论哪两部电影和HP1最相似，哪两部电影和和SW1最相似；

计算预测的电影评级与真实评级的平方误差。

计算其它电影与HP1的平方误差如下：

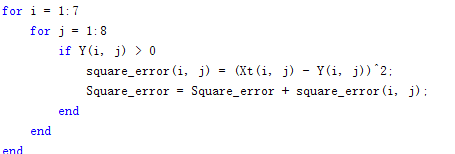


看出HP2，HP3与HP1最相似；

计算其他电影与SW1的平方误差如下：



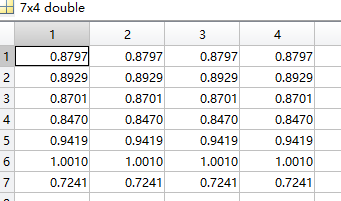
看出SW2，SW3与SW1最相似。



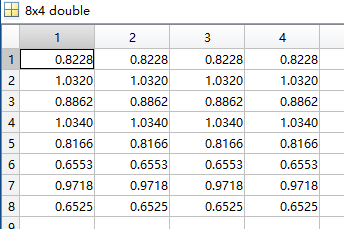
1. 试使用一个非零常数对协同过滤算法中的变量进行初始化（即更改课件第19页Collaborative ﬁltering algorithm的第一步为，其中为非零实数，为所有元素都是1的维列向量；使用（1）中相同的参数，分别计算描述电影特征的矩阵和预测用户评级的模型参数矩阵（所得结果保留小数点后4位），并计算预测电影评级的效用矩阵即（保留小数点后1位）和预测的电影评级与真实评级的平方误差，即；与（1）和（2）中的结果比较，讨论此初始化方法的问题。

答：将X和θ都初始化为全1矩阵，计算得到的结果矩阵如下：

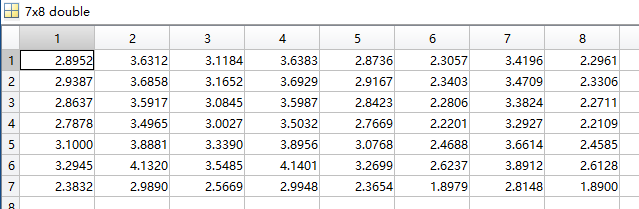
X：



Θ：



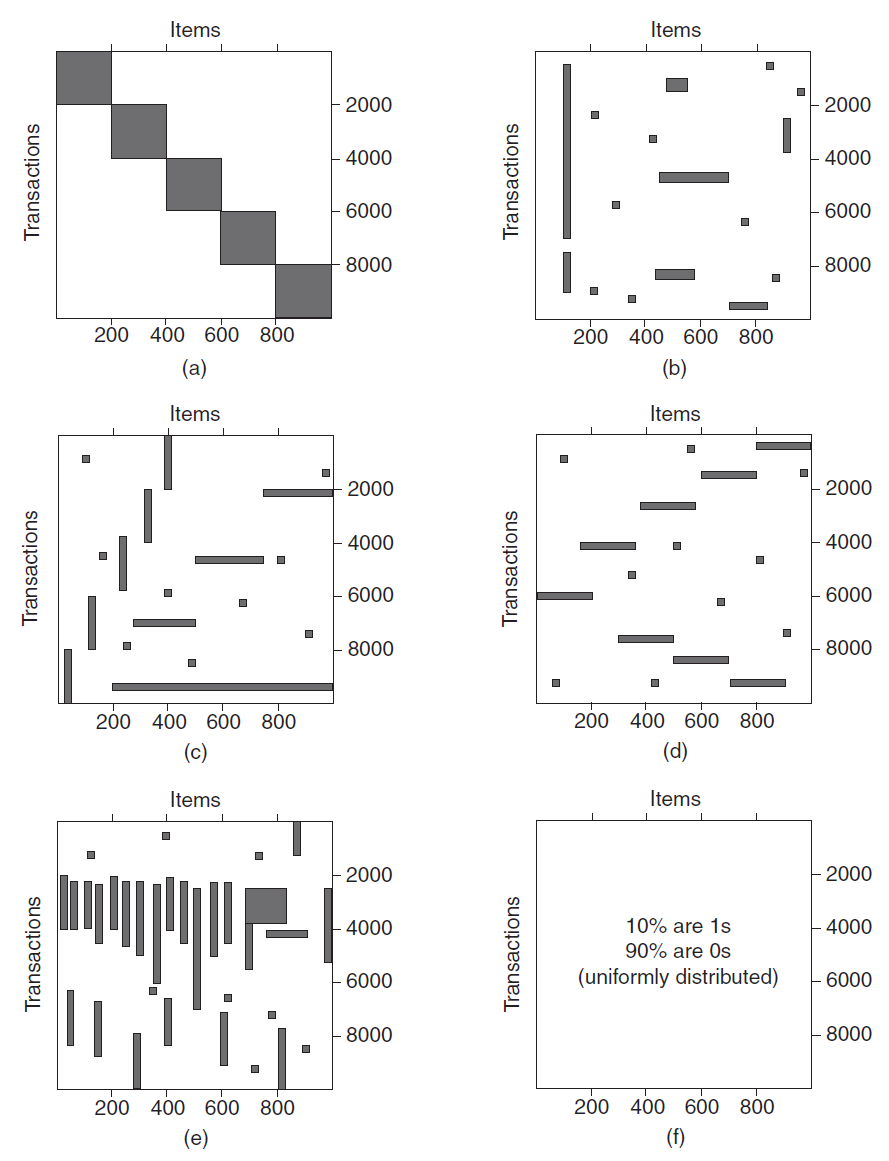
Xθ：



预测的电影评级与真实评级的平方误差为。

与（1）（2）中的结果相比，发现将参数初始化为同一个值会导致很严重的问题，算法的效果变差。这是因为参数都相同在算法学习中难以达到梯度下降的优化效果，而随机初始化通过打破对称性来确保X和θ在算法执行中保持不相同。

1. **关联规则**



上面六个图（a）到（f）中的每一个图包含1000个商品和10000个交易的记录。灰色位置表示存在商品交易，而白色表示不存在商品交易。我们使用Apriori算法提取频繁项集，并设定频繁项集的最小支持度为10%，即minsup=10%（即频繁项集包含在至少1000个交易中）。

根据上图回答以下问题：

1. 哪一个或几个数据集的频繁项集数目最多？哪一个或几个数据集的频繁项集数目最少？

a数据集的频繁项集数目最多；d数据集的频繁项集数目最少。

1. 哪一个或几个数据集的频繁项集长度最长（即包含最多商品）？

a数据集的频繁项集长度最长。

1. 哪一个或几个数据集的频繁项集有最高的最大支持度（highest maximum support）？

b数据集的频繁项集有最高的最大支持度。

1. 哪一个或几个数据集的频繁项集有最大的支持度范围（例如频繁项集的支持度范围可以从小于20%变化到大于70%）？

b数据集的频繁项集有最大的支持度范围。