Author: Pei-Zhen Li

Answers to Problem Set #3
Data Mining, Spring 2018

1 主成分分析（Principal Component Analysis）

1. **Sol:** $k = 5$时，样本均值μ对应的图像：

$Ureduce$前5个列向量所对应图像：

(a) (b) (c) (d) (e)

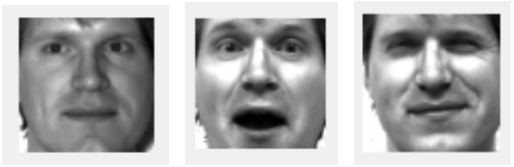
2. **Sol:** 协方差矩阵前5大特征向量对应图像：

(f) (g) (h) (i) (j)

运行时间为79.1872秒，（1）中对数据矩阵使用svd函数运行时间为0.2539秒

3. **Sol:**
$k = 10$的保留方差比例：0.7281，$k = 100$的保留方差比例：0.9824

前三张原图：



(k) (l) (m)

```
%recovery
 k=10; %k=100;
 Uk = Ur(:,1:k);
 Z = Uk'*X;
 Xre = Uk*Z;
 for i=1:m
  Xre(:,i) = Xre(:,i)+miu;
 end
```

$k = 10$对应的前三张图像：

(n) (o) (p)

$k = 100$对应的前三张图像：

(q) (r) (s)

$k = 100$时的图像与原图更接近，恢复效果更好。

2 推荐系统（Recommender System）

1. **Sol:**

Figure 1: Feature matrix X:

7x4 double

	1	2	3	4
1	0.0399	0.2812	0.6936	1.9661
2	0.8780	-0.1156	0.5789	2.2469
3	0.1631	0.1901	1.2365	1.8481
4	0.4403	0.3254	2.3329	0.2270
5	0.6337	2.1420	0.3675	0.0753
6	0.6575	2.0971	0.2964	0.1454
7	1.6367	1.8001	0.7353	-0.0264

Figure 2: Parameter matrix θ:

8x4 double

	1	2	3	4
1	0.8628	-0.1483	1.8378	1.3859
2	0.5103	-0.0125	0.3311	1.9039
3	0.7536	2.0401	0.4265	-0.0132
4	0.7399	1.3388	1.8120	-0.2338
5	0.6350	2.0491	0.2846	0.1020
6	1.4385	1.4449	-0.0543	0.2917
7	0.0295	0.3180	0.5967	2.2480
8	0.0460	0.2058	1.1615	1.3190

Figure 3: Predicted utility matrix $X\theta^{'}$:

	A	B	C	D	E	F	G	H
HP1	4.0	4.0	0.9	1.2	1.0	1.0	4.9	3.5
HP2	5.0	4.9	0.6	1.0	0.7	1.7	5.4	3.7
HP3	5.0	4.0	1.0	2.2	1.0	1.0	5.0	3.9
TW	4.9	1.4	2.0	4.9	1.6	1.0	2.0	3.1
SW1	1.0	0.6	5.0	4.0	4.9	4.0	1.1	1.0
SW2	1.0	0.7	4.9	3.8	4.8	4.0	1.2	1.0
SW3	2.5	1.0	5.2	5.0	4.9	4.9	1.0	1.3

2. **Sol:**

- **平方误差**$SE = 0.0648$

```
%Square error at each iteration step
for i=1:7
 for j=1:8
  if Y(i,j)>0
   Err(iter) = Err(iter)+(predictY(i,j)-Y(i,j))^2;
  end
 end
end
```

- 利用欧式距离计算特征向量之间的距离得到矩阵（如下），可知与HP1最相似的两部电影为HP3和HP2;与SW1最相似的两部电影为SW3和SW2.

Figure 4: Distance matrix of the movies:

	1	2	3	4	5	6	7
1	0	0.9756	0.5764	2.4237	2.7380	2.6743	2.9714
2	0.9756	0	1.0936	2.7464	3.1491	3.0726	3.0722
3	0.5764	1.0936	0	1.9812	2.8158	2.7684	2.9204
4	2.4237	2.7464	1.9812	0	2.6876	2.7093	2.4945
5	2.7380	3.1491	2.8158	2.6876	0	0.1121	1.1263
6	2.6743	3.0726	2.7684	2.7093	0.1121	0	1.1266
7	2.9714	3.0722	2.9204	2.4945	1.1263	1.1266	0

3. **Sol:**

Figure 5: Feature matrix X:

7x4 double

	1	2	3	4
1	0.9157	0.9157	0.9157	0.9157
2	0.9152	0.9152	0.9152	0.9152
3	0.9046	0.9046	0.9046	0.9046
4	0.8782	0.8782	0.8782	0.8782
5	0.9830	0.9830	0.9830	0.9830
6	1.0434	1.0434	1.0434	1.0434
7	0.7442	0.7442	0.7442	0.7442

Figure 6: Parameter matrix θ:

8x4 double

	1	2	3	4
1	0.8093	0.8093	0.8093	0.8093
2	1.0251	1.0251	1.0251	1.0251
3	0.8716	0.8716	0.8716	0.8716
4	1.0272	1.0272	1.0272	1.0272
5	0.7967	0.7967	0.7967	0.7967
6	0.6466	0.6466	0.6466	0.6466
7	0.9646	0.9646	0.9646	0.9646
8	0.6357	0.6357	0.6357	0.6357

Figure 7: Predicted utility matrix $X\theta^{'}$:

	A	B	C	D	E	F	G	H
HP1	3.0	3.8	3.2	3.8	2.9	2.4	3.5	2.3
HP2	3.0	3.8	3.2	3.8	2.9	2.4	3.5	2.3
HP3	2.9	3.7	3.2	3.7	2.9	2.3	3.5	2.3
TW	2.8	3.6	3.1	3.6	2.8	2.3	3.4	2.2
SW1	3.2	4.0	3.4	4.0	3.1	2.5	3.8	2.5
SW2	3.4	4.3	3.6	4.3	3.3	2.7	4.0	2.7
SW3	2.4	3.1	2.6	3.1	2.4	1.9	2.9	1.9

使用相同非零常数$c = 0.5$初始化，**平方误差**$SE = 88.2260$

3 关联规则（Association Rule）

1. **Sol:** （e）的频繁项集数目最多，（d）的频繁项集数目最少
2. **Sol:** （e）的频繁项集长度最长
3. **Sol:** （b）的频繁项集具有最高的最大支持度
4. **Sol:** （b）的频繁项集有最大的支持度范围