尹浚宇 161130118 计算机科学与技术 本科生

## Problem 1

**(a)** 已知题目中的矩阵$X$为非奇异实方阵，所以$X$的所有奇异值均不为$0$. 不妨令$X$的奇异值分解为$X = U\Sigma V^T$，因为$U$和$V$为正交阵，于是有$X^{-1} = V\Sigma^{-1}U^T$，不难看出$V\Sigma^{-1}U^T$即为$X^{-1}$的奇异值分解. 根据矩阵2范数的定义，有$\|X\|_2 = \sigma_{max}$，其中$\sigma_{max}$为矩阵$X$的最大奇异值. 于是得到$\kappa_2(X) = \|X\|_2\|X^{-1}\|_2 = \sigma_1/\sigma_n$.

**(b)** 由题意，假设$\boldsymbol{b}$受微小扰动变为$\boldsymbol{b}+\Delta\boldsymbol{b}$，方程的解随之变为$\boldsymbol{x}+\Delta\boldsymbol{x}$. 考虑相对误差$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$.
根据$A\boldsymbol{x} = \boldsymbol{b}$和$A(\boldsymbol{x}+\Delta\boldsymbol{x}) = \boldsymbol{b}+\Delta\boldsymbol{b}$，有$A\Delta\boldsymbol{x} = \Delta\boldsymbol{b}$，$\Delta\boldsymbol{x} = A^{-1}\Delta\boldsymbol{b}$. 根据范数的相容性，有

$$\|\Delta\boldsymbol{x}\| = \|A^{-1}\Delta\boldsymbol{b}\| \le \|A^{-1}\|\|\Delta\boldsymbol{b}\| \tag{1}$$

$$\|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \le \|A\|\|\boldsymbol{x}\| \tag{2}$$

$$\|\Delta\boldsymbol{b}\| = \|A\Delta\boldsymbol{x}\| \le \|A\|\|\Delta\boldsymbol{x}\| \tag{3}$$

$$\|\boldsymbol{x}\| = \|A^{-1}\boldsymbol{b}\| \le \|A^{-1}\|\|\boldsymbol{b}\| \tag{4}$$

根据(1)(2)，得到$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \frac{\|A^{-1}\Delta\boldsymbol{b}\|}{\|\boldsymbol{x}\|} \le \frac{\|A^{-1}\|\|\Delta\boldsymbol{b}\|}{\frac{\|\boldsymbol{b}\|}{\|A\|}} = \|A^{-1}\|\|A\|\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} = \kappa(A)\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$；根据(3)(4)，得到$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \ge \frac{\|\Delta\boldsymbol{b}\|}{\|A\|\|\boldsymbol{x}\|} \ge \frac{\|\Delta\boldsymbol{b}\|}{\|A\|\|A^{-1}\|\|\boldsymbol{b}\|} = \kappa(A)^{-1}\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$. 综上有$\kappa(A)^{-1}\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \le \frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \kappa(A)\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$. 可以看出，在外部因素$\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$固定时，$\Delta\boldsymbol{b}$引起的误差由线性系统的内部因素$\kappa(A)$决定. 如果条件数过大，则会导致解$\boldsymbol{x}$的数值稳定性较差，即$\boldsymbol{b}$的微小改变就可引起解$\boldsymbol{x}$的剧烈变化. 考虑极端情况，当$A$奇异时，条件数为无穷，这时即使不改变$\boldsymbol{b}$，$\boldsymbol{x}$也可以改变，所以病态矩阵是坏的.

**(c)** 对于任意正交矩阵$A$，有$A^T = A^{-1}$成立，则属于$A$的奇异值全为$1$，所以$\kappa_2(A) = 1$. 同时由(a)易得对于任意矩阵$X$，$\kappa_2(X)$的下界为$1$，所以正交阵拥有最小的条件数，是良态的.

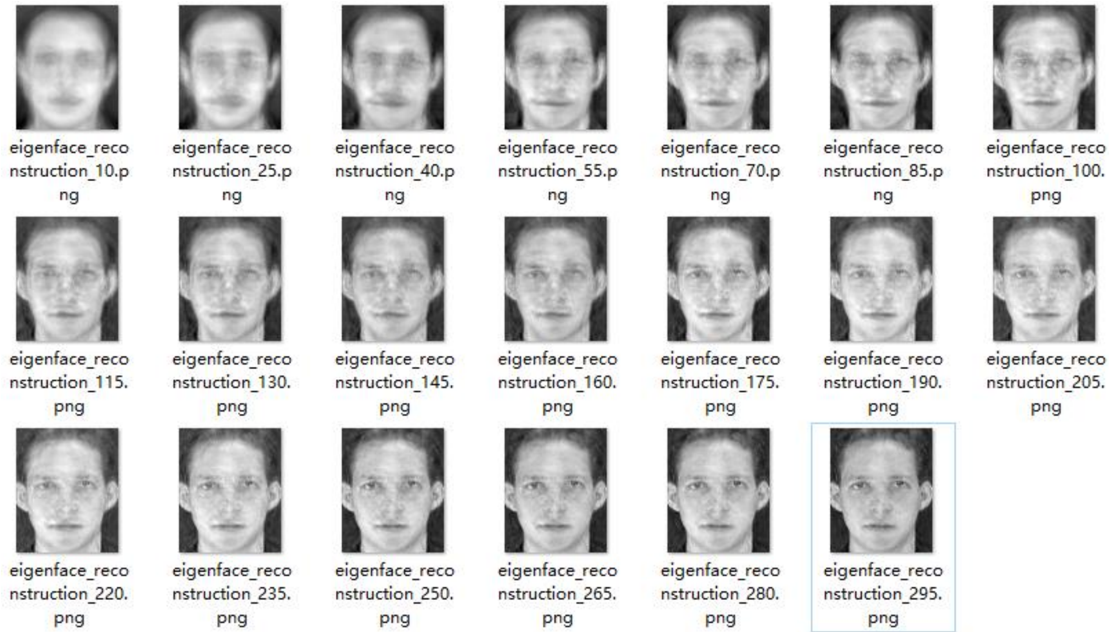
## Problem 2

(c) 结果: PCA 预测错误, FLD 预测正确.
分析: PCA 和 FLD 的本质都是对数据进行降维. 区别在于, PCA 是无监督的降维方法; FLD 是有监督的降维方法.

PCA 降维的出发点是最大化方差，其目的是去掉原始数据冗余的维度，并一定程度上减少噪声影响; FLD 降维的出发点是提取出对分类最有用的维度，即选择一个最佳的投影方向，使得投影后相同类别的数据分布紧凑，不同类别的数据尽量相互远离. 所以可以看到在实验中, PCA 人脸识别给出了错误的结果，而 FLD 给出了正确的结果. 同时在实验中, PCA 给出的特征向量具有人脸的形状，而 FLD 给出的特征向量很难看出人脸的形状，这也符合对这两种方法的分析. 同时根据数学推导, FLD 降维最多降到 C-1 维(C 是训练样本类别数量), 而 PCA 降维可以自行选择阈值来得到不同的维度.

当然, PCA 本身并不是一种人脸识别的方法，在 tutorial 中 PCA 人脸识别是通过把数据都投影到降维后的空间再使用最近邻方法来预测，这显然不会带来很好的结果. 在实际的人脸识别中，我觉得可以对数据先进行 PCA 再进行 FLD，最后使用 SVM 等方法进行训练，应该会达到不错的效果.

(d) 可以观察到至少使用 295 张 eigenfaces 来重构图像才和原图视觉上没有差别.




## Problem 3

(b)

i. Accuracy = 66.925% (2677/4000) (classification)

ii. Accuracy = 96.15% (3846/4000) (classification)

iii. Accuracy = 95.675% (3827/4000) (classification)

iv. Accuracy = 70.475% (2819/4000) (classification)

v. Accuracy = 96.525% (3861/4000) (classification)

在 SVM 方法中, 数据归一化和超参数的选择对于准确率有显著影响. 可以看到如果超参数设置不当, 复杂模型(RBF)的效果甚至不如简单的模型(线性核). 另外可以看到使用同样的模型和超参数, 数据归一化与否对结果有着巨大的影响. 最后, 对于超参数的设置可以使用交叉验证的方法, 比起靠直觉盲目选择效果更好且更有实践意义.

(c) 使用数据集: a1a, 类别+1: 395 个; 类别-1: 1210 个

不使用-wi: Accuracy = 83.5864% (25875/30956) (classification)

使用-wi: Accuracy = 75.6848% (23429/30956) (classification) (正类权重 3, 负类 1)

该情况下平衡后在测试集上的准确率反而下降了, 经过观察后我认为是测试集的正负类比例也大致为 3 比 1, 同时训练集上不同类样本的比例还不够悬殊造成的. 我认为在其他类别比例更极端的场景下, 该设置应该是可以提高准确率的.

## Problem 4

(a) 因为$p_1(x)$要满足 p.d.f 的性质, 所以有$\int_{-\infty}^{+\infty} p_1(x)\,dx = \int_{x_m}^{+\infty} p_1(x)\,dx = 1$. 即

$$\int_{x_m}^{+\infty} \frac{c_1}{x^{\alpha+1}}\,dx = \frac{c_1}{\alpha {x_m}^{\alpha}} = 1 \;\Rightarrow\; c_1 = \alpha {x_m}^{\alpha} \;\Rightarrow\; p_1(x) = \frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}} [\![x \geq x_m]\!] \sim Pareto(x_m, \alpha)$$

(b) $\ell(x_m, \alpha) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{\alpha {x_m}^{\alpha}}{{x_i}^{\alpha+1}} [\![x_i \geq x_m]\!] = \begin{cases} \frac{\alpha^n {x_m}^{n\alpha}}{(\prod_{i=1}^{n} x_i)^{\alpha+1}}, \forall i, x_i \geq x_m \\ 0 \qquad\qquad\quad , otherwise \end{cases}$

$$\ell\ell(x_m,\alpha)=\begin{cases} nln\alpha+n\alpha lnx_m-(\alpha+1)\sum_{i=1}^{n}lnx_i\,, \forall i, x_i\geq x_m \\ -\infty \qquad\qquad\qquad\qquad\qquad\qquad\quad , otherwise\end{cases}$$

不难注意到$\ell\ell(x_m,\alpha)$是关于$x_m$的单调递增函数，根据限制$\forall i, x_i \geq x_m$，于是有$\widehat{x_m}=\min\{x_1,\cdots,x_n\}$.

$$\frac{\partial\ell\ell(x_m,\alpha)}{\partial\alpha}=\frac{n}{\alpha}+nlnx_m-\sum_{i=1}^{n}lnx_i=0 \;\Rightarrow\; \hat{\alpha}=\frac{n}{\sum_{i=1}^{n}lnx_i-nln\widehat{x_m}}=\frac{n}{\sum_{i=1}^{n}ln\frac{x_i}{\widehat{x_m}}}$$

(c) $p(\mathcal{D}|\theta)=\prod_{i=1}^{n}p(x_i|\theta)=\prod_{i=1}^{n}\frac{1}{\theta}[\![0\leq x_i\leq\theta]\!]=\frac{1}{\theta^n}[\![\theta\geq\max\{x_1,\cdots,x_n\}]\!]$

$$p(\theta|x_m,k)=\frac{kx_m^k}{\theta^{k+1}}[\![\theta\geq x_m]\!]$$

$$p(\theta|\mathcal{D})=\frac{p(\mathcal{D}|\theta)p(\theta|x_m,k)}{\int p(\mathcal{D}|\theta)p(\theta|x_m,k)d\theta}=\frac{\frac{kx_m^k}{\theta^{n+k+1}}[\![\theta\geq\max\{x_1,\cdots,x_n\}]\!][\![\theta\geq x_m]\!]}{\int_{-\infty}^{+\infty}\frac{kx_m^k}{\theta^{n+k+1}}[\![\theta\geq\max\{x_1,\cdots,x_n\}]\!][\![\theta\geq x_m]\!]d\theta}$$

$$=\frac{\frac{[\![\theta\geq\max\{x_1,\cdots,x_n,x_m\}]\!]}{\theta^{n+k+1}}}{\int_{-\infty}^{+\infty}\frac{[\![\theta\geq\max\{x_1,\cdots,x_n,x_m\}]\!]}{\theta^{n+k+1}}d\theta}=\frac{\frac{[\![\theta\geq\max\{x_1,\cdots,x_n,x_m\}]\!]}{\theta^{n+k+1}}}{\int_{\max\{x_1,\cdots,x_n,x_m\}}^{+\infty}\frac{1}{\theta^{n+k+1}}d\theta}$$

$$=\frac{(n+k)(\max\{x_1,\cdots,x_n,x_m\})^{n+k}}{\theta^{n+k+1}}[\![\theta\geq\max\{x_1,\cdots,x_n,x_m\}]\!]$$

于是有$p(\theta|\mathcal{D})\sim Pareto(\max\{x_1,\cdots,x_n,x_m\},n+k)$，证毕.

## Problem 5

(b) Accuracy = 85.52% (8552/10000)

(c) Accuracy = 86.6% (8660/10000)

(d) 对数据开根号缩小了数据的范围(从 0~255 至 0~16)，同时降低了数据间的比例差异和绝对差异，等价于某种程度上的数据归一化，所以准确率有了提升.

## Problem 6

(a) 对于任意的$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathcal{X}$，$d$必须满足:

$$\begin{cases}1.\ d(\boldsymbol{x},\boldsymbol{y})\geq 0 & (\text{非负性})\\ 2.\ d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x}) & (\text{对称性})\\ 3.\ d(\boldsymbol{x},\boldsymbol{y})=0\Leftrightarrow\boldsymbol{x}=\boldsymbol{y} & (\text{同一性})\\ 4.\ d(\boldsymbol{x},\boldsymbol{z})\leq d(\boldsymbol{x},\boldsymbol{y})+d(\boldsymbol{y},\boldsymbol{z}) & (\text{三角不等式})\end{cases}$$

(b) KL 散度不是一个合法的距离度量.

$$KL(A\|B)=\frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{1}{4}}+\frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{3}{4}}\approx 0.208;\ KL(B\|A)=\frac{1}{4}\log_2\frac{\frac{1}{4}}{\frac{1}{2}}+\frac{3}{4}\log_2\frac{\frac{3}{4}}{\frac{1}{2}}\approx 0.189;$$

$$KL(A\|C)=\frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{1}{8}}+\frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{7}{8}}\approx 0.596;\ KL(C\|A)=\frac{1}{8}\log_2\frac{\frac{1}{8}}{\frac{1}{2}}+\frac{7}{8}\log_2\frac{\frac{7}{8}}{\frac{1}{2}}\approx 0.456;$$

$$KL(B\|C)=\frac{1}{4}\log_2\frac{\frac{1}{4}}{\frac{1}{8}}+\frac{3}{4}\log_2\frac{\frac{3}{4}}{\frac{7}{8}}\approx 0.083;\ KL(C\|B)=\frac{1}{8}\log_2\frac{\frac{1}{8}}{\frac{1}{4}}+\frac{7}{8}\log_2\frac{\frac{7}{8}}{\frac{3}{4}}\approx 0.070.$$

KL 散度满足性质 1：易见上述结果均大于 0.

KL 散度不满足性质 2：易见上述结果均不满足对称性.

KL 散度满足性质 3：易得$KL(A\|A),KL(B\|B),KL(C\|C)$均为 0，且上述结果均不为 0.

KL 散度不满足性质 4：有$KL(A\|C)\approx 0.596>KL(A\|B)+KL(B\|C)\approx 0.291$

(c)

```
> KL(A,B)
[1] 0.2075187
> KL(B,A)
[1] 0.1887219
> KL(A,C)
[1] 0.5963225
> KL(C,A)
[1] 0.4564356
> KL(B,C)
[1] 0.08320568
> KL(C,B)
[1] 0.06959337
```

## Problem 7

不妨令随机变量$Y$服从参数为$\frac{1}{\mu}$的指数分布，其 p.d.f 为$p(x)$. 利用 KL 散度的非负性，有

$$KL(p\|q)=\int p(x)\,ln\frac{p(x)}{q(x)}dx=-h(X)-\int p(x)\,lnq(x)dx\geq 0 \tag{7.1}$$

$$-\int p(x)\,lnq(x)dx=-\int p(x)\,ln\frac{1}{\mu}e^{-\frac{1}{\mu}x}dx=ln\mu\int p(x)\,dx+\frac{1}{\mu}\int p(x)\,xdx$$

由限制条件，有$\int p(x)\,dx=1$，$\int p(x)\,xdx=\mathbb{E}[X]=\mu$，所以有$-\int p(x)\,lnq(x)dx=1+ln\mu$.

查表得指数分布的信息熵为$1-ln\lambda$，这里即为$1+ln\mu$. 所以有$-\int p(x)\,lnq(x)dx=h(Y)$.

将上述结果代入式(7.1)，得到$-h(X)+h(Y)\geq 0$，即$h(Y)\geq h(X)$，证毕.