模式识别作业二

201300066 麻超 人工智能学院

习题3.2

a

在 k-均值聚类中，通过最小化簇内平方误差来定义簇的质心。对于每个簇 $\mathcal{C}_i$，其质心 μ_i 是所有数据点到该质心的距离平方和最小的点，即：

$$\mu_i = \arg\min_y \sum_{x \in \mathcal{C}_i} ||x-y||^2$$

然后，将每个数据点 x_j 分配到与其最近的质心 μ_i 所在的簇中：

$$\gamma_{ij} = \begin{cases} 1, & \text{if j belongs to cluster i} \\ 0, & \text{otherwise} \end{cases}$$

最小化簇内平方误差等价于最小化以下目标函数：

$$\sum_{i=1}^{K} \sum_{x \in \mathcal{C}_i} ||x-\mu_i||^2 = \sum_{i=1}^{K} \sum_{j=1}^{M} \gamma_{ij} ||x_j - \mu_i||^2$$

其中，γ_{ij} 表示将数据点 x_j 分配到簇 $\mathcal{C}_i$ 的指示变量。

因此，优化公式 $\arg\min\limits_{\gamma_{ij},\mu_i} \sum_{i=1}^{K} \sum_{j=1}^{M} \gamma_{ij} ||x_j - \mu_i||^2$ 对 k-均值聚类的目标进行了形式化。

b

在第一步中，损失函数表现为$\sum_{i=1}^{K} \sum_{j=1}^{M} \gamma_{ij} ||x_j - \mu_i||^2$.要找到一个$\gamma_{ij}$使得该式最小，对其求偏导并使其得到0，可以得到：

$$\gamma_{ij} = \begin{cases} 1, & i = \arg\min\limits_k ||x_j - \mu_k||^2 \\ 0, & \text{otherwise} \end{cases}$$

也就是说，对于每个样本x_j，我们找到最近的聚类中心μ_i，将γ_{ij}设置为1，其余设置为0。

接着，对于固定的γ_{ij}，我们需要找到μ_i使得上式最小化，将上式对μ_i求偏导并使其得到0，得到:

$$\mu_i = \frac{\sum_{j=1}^{M} \gamma_{ij} x_j}{\sum_{j=1}^{M} \gamma_{ij}}$$

也就是说，对于每个聚类中心μ_i，我们计算出属于它的所有样本x_j的均值，作为新的聚类中心。

通过这两个更新规则的迭代，Lloyd算法可以找到k-均值的解。

c

假设 $C = C_1, C_2, \cdots, C_K$ 表示数据集 X 被分成的 K 个簇，μ_i 表示簇 C_i 的中心点，$J(C, \mu)$ 表示当前簇分配和中心点的总误差平方和。在Lloyd算法中，我们迭代更新 γ_{ij} 和 μ_i，直到收敛。

在每次更新 γ_{ij} 时，我们都可以看作是在求解一个最小化 $J(C, \mu)$ 的问题，其中 μ 是固定的，而 C 是待求解的。

在每次更新 μ_i 时，我们都可以看作是在求解一个最小化 $J(C, \mu)$ 的问题，其中 C 是固定的，而 μ 是待求解的。

因此，在Lloyd算法中，每一步迭代都会使 $J(C, \mu)$ 减小，即 $J(C, \mu^{(t+1)}) \leq J(C, \mu^{(t)})$，其中 $\mu^{(t)}$ 表示第 t 次迭代后得到的中心点。因为 $J(C, \mu)$ 是非负的，所以我们可以得到 $J(C, \mu^{(t)})$ 构成一个单调递减的数列。另外，由于数据集是有限的，因此簇的数量也是有限的，所以迭代次数是有限的。所以，Lloyd算法必然会收敛到一个局部最小值。

习题4.2

a

线性回归任务即找到最优的β，使代价最小化，据此可得平方误差为$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$.则线性回归任务可以表示为如下的优化问题:

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

b

用X和y重写可以表示为:

$$\arg\min_{\beta} (y - X\beta)^T (y - X\beta)$$

c

令平方误差
$J(\beta) = (y - X\beta)^T (y - X\beta) \ = (y^T - \beta^T X^T)(y - X\beta) \ = (y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta)$

对上式取β的偏导，可以得到

$\frac{\partial J(\beta)}{\partial \beta} = (2X^T X\beta - X^T y - y^T X) \ = (2X^T X\beta - 2X^T y) \qquad \text{since}(X^T y = y^T X)$

令上式为0，且由于$X^T X$可逆，可以得到 $\beta^* = (X^T X)^{-1} X^T y$

d

此时有$d > n$,且矩阵X是一个$n \times d$矩阵，则可以得到矩阵X的秩$rank(X) \leq n < d$.

所以$rank(X^T X) = rank(X) < d$

但矩阵$X^T X$是一个$d \times d$的矩阵，故其必然不满秩，且不可逆

e

该正则项会对模型产生的影响有：提高模型泛化能力，防止过拟合现象的发生；存在共线性时，减少参数方差，提高模型稳定性和精度；实际应用中，如果样本量较小，正则项可以提高模型的估计准确性。

f

岭回归的优化问题是求

$$\arg\min_{\beta}((y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta)$$

同样求导并令导数为0，可得

$$\hat{\beta}=(X^TX+\lambda I)^{-1}X^Ty$$

g

当 X^TX 不可逆时，$(X^TX+\lambda I)$ 是可逆的，因为正则化项 $\lambda||\beta||^2$ 强制使得 $(X^TX+\lambda I)$ 始终是满秩的，从而可以求解 β 的值。当 λ 取较大值时，正则化项的作用更强，模型更加稳定，但是预测的偏差会变大，而当 λ 取较小值时，模型更加灵活，但是预测的方差会变大。

h

$\lambda=0$时，岭回归的解就是普通线性回归的解，也就是

$$\hat{\beta}=(X^TX)^{-1}X^Ty$$

当$\lambda=\infty$时，此时正则化项的影响最大，只能解出$\hat{\beta}=0.$

i

可以。我们可以引入一个新的超参数α，其中$\lambda=\alpha/n, n$是训练集中样本数量。我们可以通过交叉验证等技术选取最佳的α值，进而重新训练模型，获得最终的β值。

习题4.5

a

下标	类别标记	得分	查准率P	查全率R	AUC-PR	AP
0	-	-	1.0000	0.0000	-	-
1	1	1.0	1.0000	0.2000	0.2000	0.2000
2	2	0.9	0.5000	0.2000	0.0000	0.0000
3	1	0.8	0.6667	0.4000	0.1167	0.1333
4	1	0.7	0.7500	0.6000	0.1417	0.1333
5	2	0.6	0.6000	0.6000	0.0000	0.1500
6	1	0.5	0.6777	0.8000	0.1267	0.0000
7	2	0.4	0.5714	0.8000	0.0000	0.1333

下标	类别标记	得分	查准率P	查全率R	AUC-PR	AP
8	2	0.3	0.5000	0.8000	0.0000	0.0000
9	1	0.2	0.5556	1.000	0.1056	0.1111
10	2	0.1	0.5000	1.000	0.0000	0.0000
-	-	-			0.6906	0.7278

b

是这样的，AUC-PR和AP的值应该逼死相似，用AP的表达式减去AUC_PR的表达式可以得到:

$$
\begin{aligned}
AP - AUC\_PR &= (r_i - r_{i-1})p_i - (r_i - r_{i-1})\frac{p_i + p_{i-1}}{2} \\
&= \frac{1}{2}(r_i - r_{i-1})(p_i - p_{i-1})
\end{aligned}
$$

所以AP每次只会比AUC-PR大一点，但二者总体上还是接近的。

c

新的AUC_PR为`0.6794`，新的AP为`0.7167`

d

程序见附件`Q3AUC_PR&AP.py`核心部分代码如下:

```
labels=[1,2,1,1,2,1,2,2,1,2]

#Question c swap line 9 and line 10
#labels=[1,2,1,1,2,1,2,2,2,1]

P = [1.0]
R = [0.0]
AUC_PR=[]
AP=[]

for i in range(len(scores)):
    P.append(Count(1,labels[:i+1])/(i+1))
    R.append(Count(1,labels[:i+1])/Count(1,labels))
    AUC_PR.append((P[i+1]+P[i])*(R[i+1]-R[i])/2)
    AP.append((R[i+1]-R[i])*P[i+1])
AUC_PR_SUM=sum(AUC_PR)
AP_SUM=sum(AP)
```

得到结果正确。

习题4.6

a

$$\begin{aligned}\mathbb{E}_{\mathrm{D}}[(y-f(x;D))^2] &= \mathbb{E}_{\mathrm{D}}[(F(x)-f(x;D)+\epsilon)^2] \\ &= \mathbb{E}_{\mathrm{D}}[(F(x)-f(x;D))^2+\epsilon^2+2(F(x)-f(x;D))\epsilon] \\ \because \mathbb{E}_{\mathrm{D}}[(F(x)-f(x;D))^2] &= (\mathbb{E}_{\mathrm{D}}[F(x)-f(x;D)])^2+Var(F(x)-f(x;D)) \\ \text{且}(\mathbb{E}_{\mathrm{D}}[F(x)-f(x;D)])^2 &= (F(x)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2 \\ Var(F(x)-f(x;D)) &= Var(-f(x;D)) = Var(f(x;D)) = \mathbb{E}_{\mathrm{D}}[(f-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2] \\ \text{综上所述，}\mathbb{E}_{\mathrm{D}}[(y-f(x;D))^2] &= (F-\mathbb{E}_{\mathrm{D}}[f_D(x;D)])^2+\mathbb{E}_{\mathrm{D}}[(f(x;D)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2]+\sigma^2\end{aligned}$$

由独立性我们可知，

$$\mathbb{E}_{\mathrm{D}}[\epsilon^2] \quad = (\mathbb{E}_{\mathrm{D}}[\epsilon])^2+Var(\epsilon)=\sigma^2$$

同时有：

$$\mathbb{E}_{\mathrm{D}}[(F(x)-f(x))\epsilon] = \mathbb{E}_{\mathrm{D}}[F(x)-f(x;D)]\mathbb{E}_{\mathrm{D}}[\epsilon]=0$$

所以有：

$$\mathbb{E}_{\mathrm{D}}[(y-f(x;D))^2] = \mathbb{E}_{\mathrm{D}}[(F(x)-f(x;D))^2]+\sigma^2$$

此时进一步展开$\mathbb{E}_{\mathrm{D}}[(y-f(x;D))^2]$，可得：

$$\mathbb{E}_{\mathrm{D}}[(F(x)-f(x;D))^2] = (\mathbb{E}_{\mathrm{D}}[F(x)-f(x;D)])^2+Var(F(x)-f(x;D))$$

由于$F(x)$与训练集D无关,是确定的，所以有$\mathbb{E}_{\mathrm{D}}[F(x)]=F(x)$，且由于$F(x)$确定，不影响方差，即$Var(F(x)-f(x;D)) = Var(-f(x;D)) = Var(f(x;D)) = \mathbb{E}_{\mathrm{D}}[(f(x;D)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2]$故偏置-方差分解为：

$$\mathbb{E}_{\mathrm{D}}[(y-f(x;D))^2] = (F-\mathbb{E}_{\mathrm{D}}[f_D(x;D)])^2+\mathbb{E}_{\mathrm{D}}[(f(x;D)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2]+\sigma^2$$

第一项为真实标记$F(x)$与所有训练集D下期望输出标记的偏差，第二项为$f(x;D)$关于训练集D的方差，第三项为噪声ϵ关于训练集D的方差。

b

$$\mathbb{E}[f] = \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} y_{nn(i)}] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[F(x_{nn(i)})+\epsilon] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[F(x_{nn(i)})】$$

c

$$\begin{aligned}\mathbb{E}_{\mathrm{D}}[y-f(x;D)] =&(F(x)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2+\mathbb{E}_{\mathrm{D}}[(f(x;D)-\mathbb{E}_{\mathrm{D}}[f(x;D)])^2]+\sigma^2 \\ =&(F(x)-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})])^2+\mathbb{E}_{\mathrm{D}}[\frac{1}{k}\sum_{i=1}^{k}y_{nn(i)}-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})])^2]+\sigma^2 \\ =&(F(x)-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})])^2+\frac{1}{k^2}\mathbb{E}_{\mathrm{D}}[\sum_{i=1}^{k}(y_{nn(i)}-\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})]))^2]+\sigma^2\end{aligned}$$

d

方差项为$\frac{1}{k^2}\mathbb{E}_{\mathrm{D}}[\sum_{i=1}^{k}(y_{nn(i)}-\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})]))^2]$

当k增加时，方差项会先减小。k增加时，在训练集上寻找的最近邻数会更多，导致方差项内$(y_{nn(i)}-\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})]])$接近于0，且$\frac{1}{k^2}$减小，导致方差项减小。

e

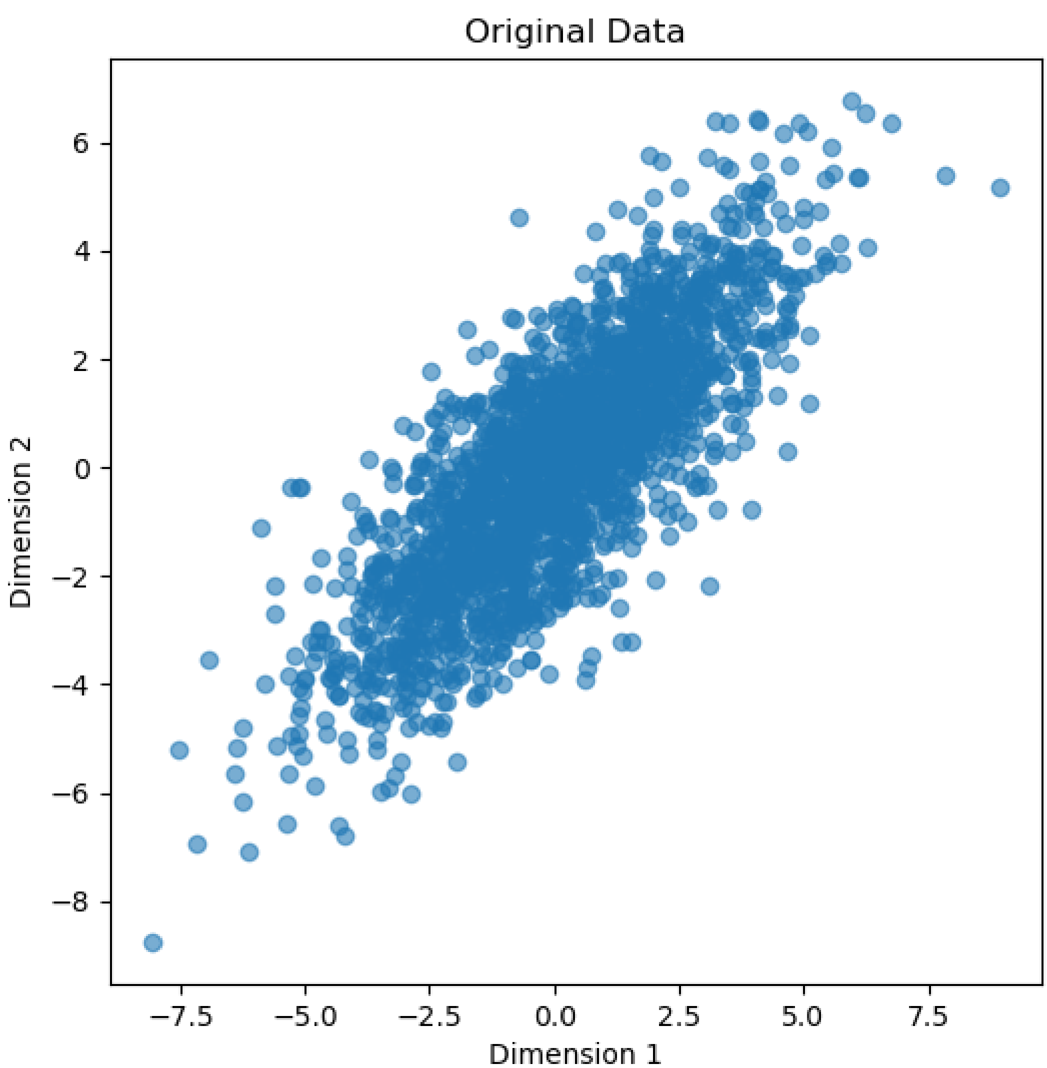
偏置的平方项为$(F(x)-\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})]])^2$.

偏置的平方项会随着k的增大而减小。当$k=n$时，此时有$E(\epsilon)=0$，所以$F(x)=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathrm{D}}[F(x_{nn(i)})]]$,偏置项为0

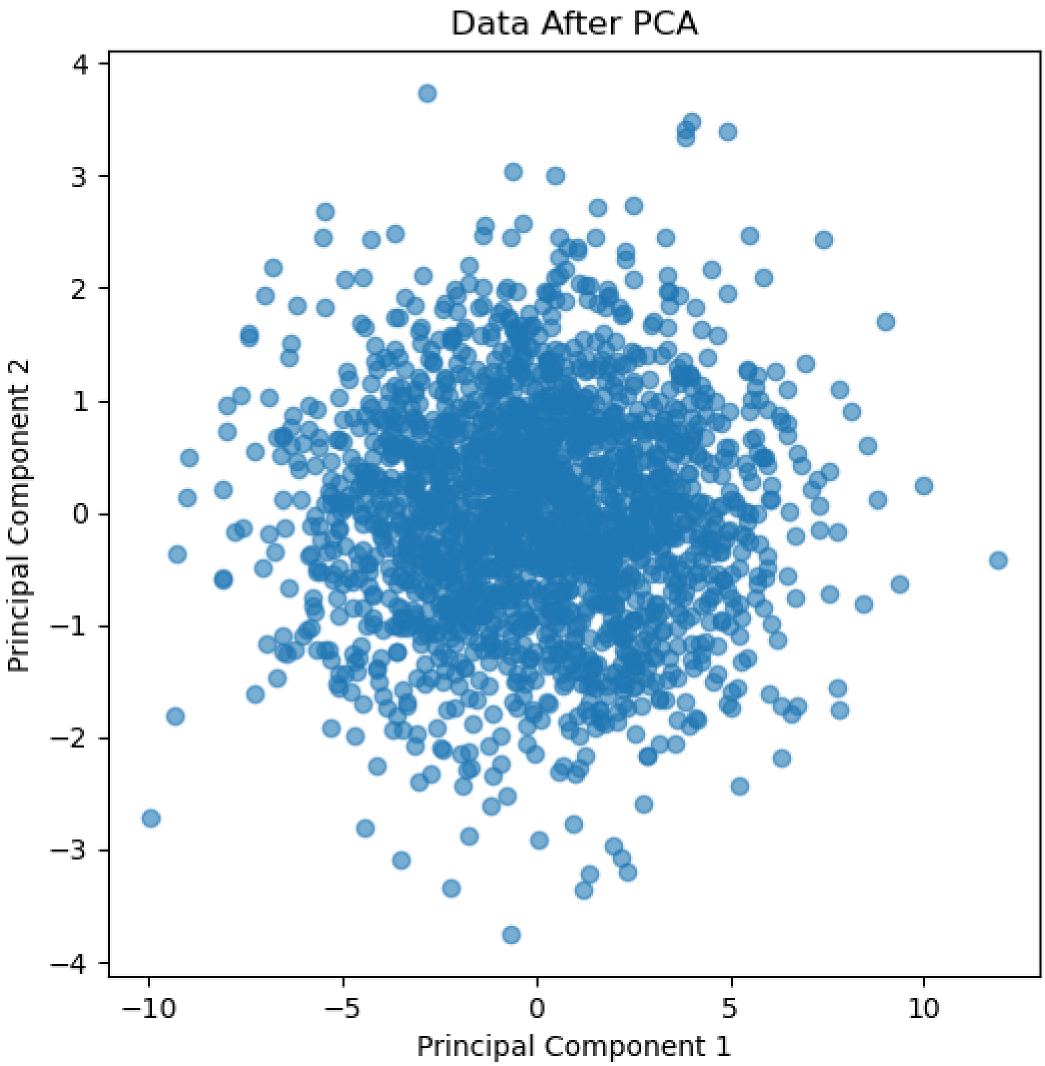
习题5.3

a

代码见附件`Q5_PCA.py`以下为运行结果

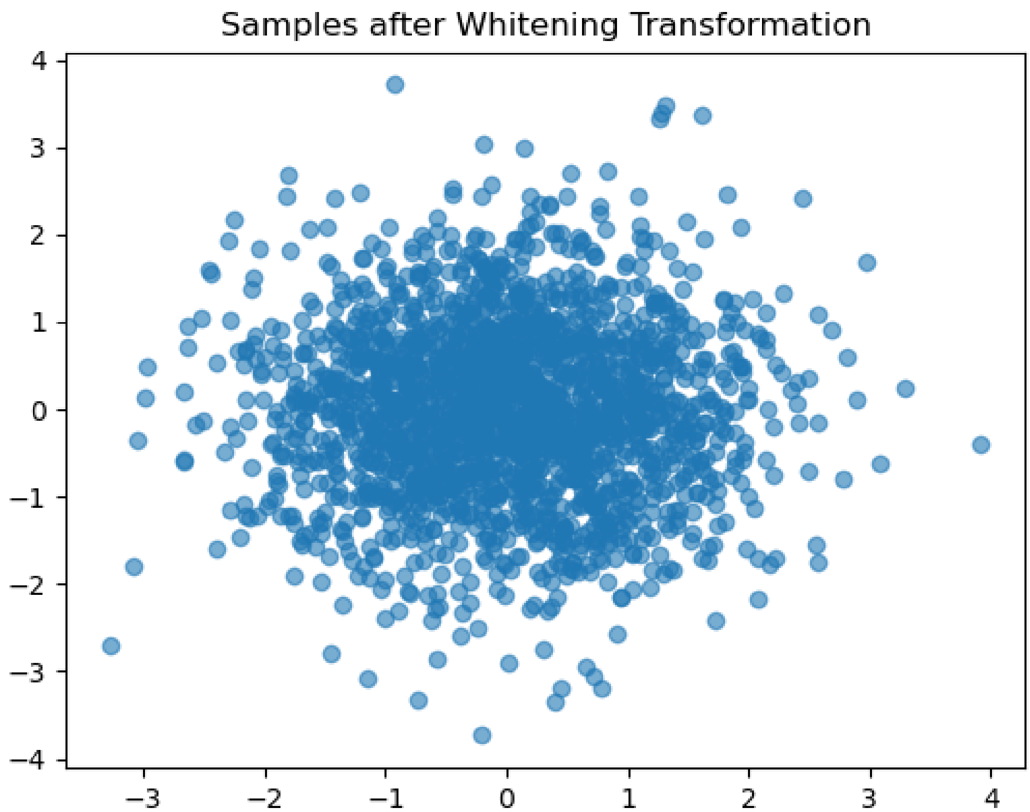


b



Data After PCA
Principal Component 2
Principal Component 1
4
3
2
1
0
−1
−2
−3
−4
−10
−5
0
5
10

c



d

在保留所有维度的情况下，PCA将原始数据变换为一个新的坐标系，其中每个坐标轴是原始数据中不同维度之间的无关方向。也就是说PCA对数据进行的是一个正交变换，这个变换是通过对数据协方差矩阵进行特征值分解得到的。

PCA的这一操作通过对数据进行旋转，可以更好地理解和描述数据，同时也能够去除数据中的冗余信息和噪声，从而使得数据更加紧凑和易于处理。此外，在某些情况下，数据的不同维度之间可能存在相关性，而PCA可以通过将数据旋转到一个新的坐标系中来减少这些相关性的影响，使得不同维度之间的独立性更加明显。

习题6.3

a

由矩阵2-范数的定义和矩阵的逆的性质，可知$||\mathrm{X}||_2 = \sigma_1, ||\mathrm{X}^{-1}||_2 = \frac{1}{\sigma_n}$

故$\kappa_2(X) = ||\mathrm{X}||_2||\mathrm{X}^{-1}||_2 = \frac{\sigma_1}{\sigma_n}$

b

当矩阵A为病态的时，意味着A的最大奇异值与最小奇异值之比很大，或者表示为其条件数$\kappa_2(A)$很大。

在这种情况下，对输入向量b或矩阵A的任何小扰动可能会导致输出解x发生较大变化。这使得求解$Ax = b$的问题在数值上不稳定且不可靠。

例如，假设我们正在求解一个线性方程组，以根据一些有噪声的测量值来估计某些参数时，病态的线性系统可能导致数据非常敏感并且产生不稳定的估计，对结果产生较大的偏差。

c

令Q为正交矩阵，即$Q^TQ = I$，I为单位矩阵。 欲证 Q良态，即需要证明 $\kappa(Q)$ 很小。

由于正交矩阵Q保留了向量的欧几里德范数，且Q的最大奇异值是1，所以有$||Q||_2 = 1$.故对于所有向量 x，有 $||Qx||_2 = ||x||_2$

另一方面，由于Q正交，故$Q^{-1} = Q^T$，且对于任何矩阵 A，恒有$||A^T||_2 = ||A||_2$.故 $||Q^{-1}||_2 = ||Q^T||_2 = ||Q||_2 = 1$。

所以根据2-范数条件，矩阵Q的条件数为$\kappa_2(Q) = ||Q||_2||Q^{-1}||_2 = 1$，所以$Q$是良态的。