1. （8 分）试阐述线性判别函数的基本概念，并说明既然有线性判别函数，为什么还需要非线性判别函数？假设有两类模式，每类包括 6 个 4 维不同的模式，且良好分布。如果它们是线性可分的，问权向量至少需要几个系数分量？假如要建立二次的多项式判别函数，又至少需要几个系数分量？（设模式的良好分布不因模式变化而改变）

i. 试阐述线性判别函数的基本概念，并说明既然有线性判别函数，为什么还需要非线性判别函数?

   a) 线性判别函数的一般函数形式是$y = w^T x$ ,其中$x$是特征的增广向量，$w$则是权重系数，一般根据$y$的取值来进行类别判定，比如 2 类问题可以定$y > 0 \rightarrow w1\,; y < 0 \rightarrow w2$。因为这个函数的几何形态往往是一条直线（或者多维下的超平面），所以称为线性判别。如果$x$是经过低维向高维投影的特征，则是广义线性判别函数。

   b) 虽然广义线性判别函数可以达到非线性判别的效果，但是随着模型复杂度的提升，往往会遇到参数爆炸的问题，采用核技巧虽然可以避开参数爆炸，但是也会遇到 kernel 形式有限和没有 kernel 是否合适的评估机制的弊端。因此如果能够基于先验知识确定一个合适的非线性判别函数，还是会避开很多问题而取得较好效果的。

   c) 包括 6 个 4 维不同的模式（样本？），则线性权向量至少多少？二次权向量至少要多少?

      i. 线性权向量至少要 5 个（d+1）

      ii. 二次权向量至少 15 个 $(4(一次项) + 4(二次项) + C_4^2(混合项) + 1(w_0))$

      iii. 公式$\frac{(n+r)!}{n!r!}$

         1. 线性 $n = 4, r = 1\,; \frac{(n+r)!}{n!r!} = 5$

         2. 二次$n = 4, r = 2\,; \frac{(n+r)!}{n!r!} = 15$

---

2. （8 分）简述 SVM 算法的原理。如果使用 SVM 做二分类问题得到如下结果，分别应该采取什么措施以取得更好的结果？并说明原因。
(1) 训练集的分类准确率 90%，验证集的分类准确率 90%，测试集的分类准确率 88%；
(2) 训练集的分类准确率 98%，验证集的分类准确率 90%，测试集的分类准确率 88%。

i. SVM 的算法的原理

   a) 一言以蔽之：最大化分类 margin 。在 soft margin 的情况下，其实是求解下面问题的最优解：

$$\min \frac{1}{2} ||w||^2 \quad + C \sum_{i=1}^{n} \varepsilon_i$$

$$st \qquad y^i(w^T x^i + b) > 1 - \varepsilon_i\,, i = 1, 2, \dots, n$$

$$\varepsilon_i \geq 0\,, i = 1, 2, \dots, n$$

使用 Lagranger 函数处理再取其对偶问题是：（得到的$\alpha_i != 0$的就是支持向量，分类面在支持向量正中间。

$$max_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y^i y^j \alpha_i \alpha_j (x^i)^T x^j$$

$$st \quad 0 \le \alpha_i \le C\,, i = 1, 2, \dots, n$$

$$\sum_{i=0}^{n} \alpha_i y^i = 0$$

ii. 训练集合，测试集合，验证集合的准确性都大约是 90% ，可以适当的增大小 C 值再训练，因为此时的模型尚未出现过拟合，同时准确率没有非常高，说明模型对错误的容忍过于宽泛，margin 宽余实际需求。

iii. 训练集 98% ，测试和验证集合都越 90% ，可以适当的减小 C 值，因为感觉已经过拟合，训练集合对错误过于严苛，margin 太小导致泛化能力差。

---

3. （8 分）请从两种角度解释主成分分析（PCA）的优化目标。

预设$W$是变换矩阵，$x$是原始特征向量，$z = W^T x$是变化后的向量。不失一般性的假定样本中心是坐标原点。

i. 最大化映射后的样本方差角度
   a) $max \sum_i z_i^T z_i = max_w (W^T X X^T W)$

ii. 最小重建误差角度
   a) $\min \sum_{i=1}^{n} (x_i - W(W^T x_i))^2$ 求最佳$W$
      i. 这个公式可以推导成最大方差公式

---

4. （8 分）请给出卷积神经网络 CNN 中卷积、Pooling、RELU 等基本层操作的含义。然后从提取特征的角度分析 CNN 与传统特征提取方法（例如 Gabor 小波滤波器）的异同。

i. 基本层操作
   a) 卷积 ： 部分特征与滤波器做矩阵乘（相乘后求和为卷积）的操作，是一种局部特征提取的手段。
   b) Pooling ： 池化，将局部特征压缩（比如 2x2 → 1）的手段。 池化是逐步扩大卷积的范围有效手段，从而使得在计算量不显著上升的情况下得到卷积也能获得更加全局的特征。
   c) ReLU ： 神经元非线性转移的一种。$y = x\ (x > 0)\ or\ y = 0\ (x \le 0)$ 。 是一种可以解决梯度消失的转移函数。不过会带来神经元死亡的问题。

ii. 异同
   a) 相同： 不同的特征之间的权值共享
   b) 不同： CNN 的权值是学习获得的，Gabor 的权重是预设的

---

5. （10 分）用线性判别函数的感知器赏罚训练算法求下列模式分类的解向量，并给出相应的判别函数。

$\omega_1$: $\{(0\ 0)^T,\ (0\ 1)^T\}$

$\omega_2$: $\{(1\ 0)^T,\ (1\ 1)^T\}$

i. 使用批处理感知器
   a) 获得规范增广矩阵

      $\{\{0,\ 0,\ 1\},\ \{0,\ 1,\ 1\},\ \{-1,\ 0,\ -1\},\ \{-1,\ -1,\ -1\}\}$

   b) 初始化向量$w = (1,1,1)$，步长 1
   c) 迭代
      i. $wx1 > 0$；$wx2 > 0$；$wx3 < 0$；$wx4 < 0$； $w = (-1,0,-1)$
      ii. $wx1 < 0$；$wx2 < 0$；$wx3 > 0$；$wx4 > 0$； $w = (-1,1,1)$
      iii. $wx1 > 0$；$wx2 > 0$；$wx3 = 0$；$wx4 < 0$； $w = (-3,0,-1)$
      iv. $wx1 < 0$；$wx2 < 0$；$wx3 > 0$；$wx4 > 0$；$w = (-3,1,1)$
      v. $wx1 > 0$；$wx2 > 0$；$wx3 > 0$；$wx4 > 0$；$done$
   d) 判别函数 $y = (-3,1)^T x + 1$； $if\ y > 0 \rightarrow w_1$； $if\ y < 0 \rightarrow w_2$

---

6. （10 分）试述 K-L 变换的基本原理，并将如下两类样本集的特征维数降到一维，同时画出样本在该空间中的位置。

$\omega_1$: $\{(-5\ -5)^T,\ (-5\ -4)^T,\ (-4\ -5)^T,\ (-5\ -6)^T,\ (-6\ -5)^T\}$

$\omega_2$: $\{(5\ 5)^T,\ (5\ 6)^T,\ (6\ 5)^T,\ (5\ 4)^T,\ (4\ 5)^T\}$，

其中假设其先验概率相等，即 $P(\omega_1)=P(\omega_2)=0.5$。

i. K-L 变换的基本原理：
   a) K-L 的关注问题是在均方误差最小的条件下获得最佳降维变换。
   b) 算法步骤是：
      i. 将特征减去均值$X = X - E[X]$
      ii. 计算协方差矩阵$C = XX^T$
      iii. $C$进行特征值分解，获得的特征向量按照特征值大小排序，取其前$k$个作为转移矩阵$W$
      iv. $W^T X$ 就是降维后的特征
ii. 对样本进行降维
   a) $E(X) = (0,0)^T$ 符合最佳 K-L 变换需求
   b) $C = X^T X = \begin{Bmatrix} 254 & 250 \\ 250 & 254 \end{Bmatrix}$
   c) $(C - \lambda I)x = 0 \rightarrow \begin{vmatrix} 254-\lambda & 250 \\ 250 & 254-\lambda \end{vmatrix} = 0$

$$\rightarrow \begin{vmatrix} 4-\lambda & -4+\lambda \\ 250 & 254-\lambda \end{vmatrix} = 0$$

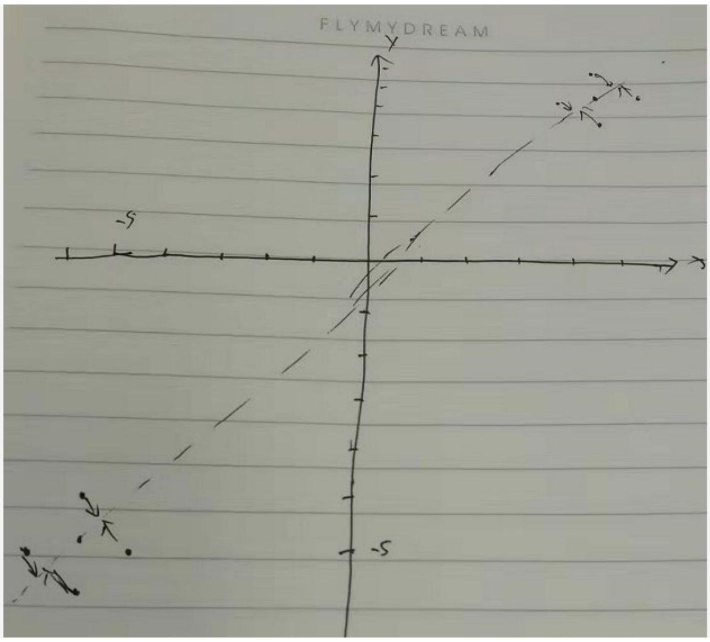
$$\rightarrow (4-\lambda)(504-\lambda) = 0$$

i. $\rightarrow \lambda_1 = 504 \rightarrow$ 特征向量$w = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$

降维： $w = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$

ii. $W^TX = \{-\frac{10}{\sqrt{2}}, -\frac{9}{\sqrt{2}}, -\frac{9}{\sqrt{2}}, -\frac{11}{\sqrt{2}}, -\frac{11}{\sqrt{2}}, \frac{10}{\sqrt{2}}, \frac{11}{\sqrt{2}}, \frac{11}{\sqrt{2}}, \frac{9}{\sqrt{2}}, \frac{9}{\sqrt{2}}\}$

d)




---

（12 分）请解释 AdaBoost 的基本思想和工作原理，写出 AdaBoost 算法

i. 基本思想：
   a) 构造强学习往往难度较大，构造弱学习器则不难，如果能构造多个弱学习器并使得他们能够互补的话，就能够组合出好性能。
   b) adaboost 采用在弱学习器 1 上失败的样本训练弱学习器 2
      i. 确保弱学习器 1 在其训练集上误差<0.5
   c) 调整样本权重，使得弱学习器 1 在样本上表现等于随机猜想。
      i. 然后用调整过权重的样本来训练弱学习器 2

ii. Adaboost 算法
   a) 给定训练集合：$(x_1, y_1), \ldots (x_n, y_n)$ 其中$y \in \{1, -1\}$ 表示类别标签
   b) 初始化样本权重$w_{1,i} = \frac{1}{N}$
   c) 迭代$m = 1 : M$
      i. 对训练样本采用权重$w_{m,i}$ 训练弱分类器$\emptyset_m(x)$
      ii. 计算当前权重下误差 $\varepsilon_m = \sum_{i=1}^{N} w_{m,i} \mathbb{I}(\emptyset_m(x_i) != y_i)$

iii. 更新权重$w_{m+1,i} = \frac{w_{m,1}\exp\left(-\alpha_m y_i \emptyset_m(x_i)\right)}{Z_m}$ 其中

1. $\alpha_m = \frac{1}{2}\log\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$
2. $Z_m$是归一化因子

d) 最终的强分类器是 $sgn\left(\sum_{i=1}^{M} \alpha_m \emptyset_m(x_i)\right)$

---

8. （12分）选择埃尔米特多项式，其前面几项的表达式为

$H_0(x)=1$, $H_1(x)=2x$, $H_2(x)=4x^2-2$,

$H_3(x)=8x^3-12x$, $H_4(x)=16x^4-48x^2+12$

试用二次埃尔米特多项式的势函数算法求解以下模式的分类问题

$\omega_1$: $\{(0\ 1)^T, (0\ -1)^T\}$

$\omega_2$: $\{(1\ 0)^T, (-1\ 0)^T\}$

i. 构造正交函数集合，根据题干要求需要二次项，因此取$H_0$和$H_2$构建即可：

$$\phi_1(\boldsymbol{x}) = H_0(x_1) * H_0(x_2) = 1$$
$$\phi_2(\boldsymbol{x}) = H_0(x_1) * H_2(x_2) = 4x_2^2 - 2$$
$$\phi_3(\boldsymbol{x}) = H_2(x_1) * H_0(x_2) = 4x_1^2 - 2$$
$$\phi_4(\boldsymbol{x}) = H_2(x_1) * H_2(x_2) = 16x_1^2x_2^2 - 8x_1^2 - 8x_2^2 + 4$$

ii. 构造核函数

$$K(\boldsymbol{x_i}, \boldsymbol{x_k}) = \sum_m \phi_m(\boldsymbol{x_i})\phi_m(\boldsymbol{x_k}) = 1$$
$$+16x_{i,1}^2x_{k,1}^2 - 8x_{i,1}^2 - 8x_{k,1}^2 + 4$$
$$+16x_{i,2}^2x_{k,2}^2 - 8x_{i,2}^2 - 8x_{k,2}^2 + 4$$
$$+\left(16x_{i,1}^2x_{i,2}^2 - 8x_{i,1}^2 - 8x_{i,2}^2 + 4\right)\left(16x_{k,1}^2x_{k,2}^2 - 8x_{k,1}^2 - 8x_{k,2}^2 + 4\right)$$

iii. 训练

迭代直到全部可以分类：

$$K_0(\boldsymbol{x}) = 0\ ;$$
$$K_{i+1}(\boldsymbol{x}) = K(\boldsymbol{x}) + K(\boldsymbol{x}, \boldsymbol{x_i})\ \boldsymbol{if}\ K(\boldsymbol{x_i}, \boldsymbol{x}) \le \boldsymbol{0}\ \&\&\ \boldsymbol{w_i} = \boldsymbol{1}$$
$$K_{i+1}(\boldsymbol{x}) = K(\boldsymbol{x}) - K(\boldsymbol{x}, \boldsymbol{x_i})\ \boldsymbol{if}\ K(\boldsymbol{x_i}, \boldsymbol{x}) \ge \boldsymbol{0}\ \&\&\ \boldsymbol{w_i} = \boldsymbol{-1}$$

---

9. （12分）已知以下关于垃圾邮件的8条标注数据，A、B为邮件的2个特征，Y为类别，其中Y=1表示该邮件为垃圾邮件，Y=0表示该邮件为正常邮件。请依此训练一个朴素贝叶斯分类器，并预测特征为“A=0，B=1”的邮件是否为垃圾邮件。

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

手动修正题干

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Y | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

$$p(y=1)=\frac{3}{8}$$

$$p(A=0)=0.25$$

$$p(B=1)=0.25$$

$$p(A=0|Y=1)=\frac{1}{3}$$

$$p(A=1|Y=1)=\frac{2}{3}$$

$$p(A=0|Y=0)=\frac{1}{5}$$

$$p(A=0|Y=0)=\frac{4}{5}$$

$$p(B=0|Y=1)=\frac{2}{3}$$

$$p(B=1|Y=1)=\frac{1}{3}$$

$$p(B=0|Y=0)=\frac{4}{5}$$

$$p(B=1|Y=0)=\frac{1}{5}$$

$$p(Y=1\mid A=0,B=1)=\frac{p(A=0,B=1\mid Y=1)p(Y=1)}{p(A=0,B=1|p(Y=1)+p(A=0,B=1|p(Y=0)}$$

朴素贝叶斯认为属性独立:

$$=\frac{p(A=0|Y=1)*p(B=1|Y=1)*p(Y=1)}{p(A=0|Y=1)*p(B=1|Y=1)*p(Y=1)+p(A=0|Y=0)*p(B=1|Y=0)*p(Y=0)}$$

$$=\frac{\frac{1}{3}*\frac{1}{3}*\frac{3}{8}}{\frac{1}{3}*\frac{1}{3}*\frac{3}{8}+\frac{1}{5}*\frac{1}{5}*\frac{5}{8}}=\frac{\frac{1}{24}}{\frac{1}{24}+\frac{1}{40}}=\frac{40}{40+24}=0.625$$

是垃圾邮件!

---

10. （12 分）假设有 3 个罐子，每个罐子里都装有红、黑两种颜色的弹珠。按照下面的方法取弹珠：开始，以概率 $\pi$ 随机选取 1 个罐子，从这个罐子以概率 B 随机取出一个弹珠，记录其颜色后，放回；然后，从当前盒子以概率 A 随机转移到下一个盒子，再从这个盒子里以概率 B 随机抽出一个球，记录其颜色，放回；如此重复 3 次，得到一个弹珠的颜色观测序列：O=（红，黑，红）。请用前向传播算法计算生成该序列的概率 $P(O|\{A,B,\pi\})$。

$$\pi=[0.4,0.4,0.2]^T \quad A=\begin{array}{c|ccc} & 罐子1 & 罐子2 & 罐子3 \\ \hline 罐子1 & 0.3 & 0.5 & 0.2 \\ 罐子2 & 0.2 & 0.3 & 0.5 \\ 罐子3 & 0.1 & 0.4 & 0.5 \end{array} \quad B=\begin{array}{c|cc} & 红 & 黑 \\ \hline 罐子1 & 0.7 & 0.3 \\ 罐子2 & 0.5 & 0.5 \\ 罐子3 & 0.4 & 0.6 \end{array}$$

i. 参考公式

$$p(y_t|x)=\frac{p(x_1,\dots x_t,y_t)p(x_{t+1},\dots x_T|y_t)}{p(x)}$$
$$\alpha(t)=p(x_1,\dots x_t,y_t)$$
$$\beta(t)=p(x_{t+1},\dots x_T|y_t)$$
$$p(x)=\sum\nolimits_{y_t}\alpha(t)\beta(t)$$
$$\alpha(t+1)=\sum\nolimits_{y_t}\alpha(t)a_t a_{t+1}p(x_{t+1}|y_{t+1})$$
$$\beta(t)=\sum\nolimits_{y_{t+1}}\beta(t+1)a_t a_{t+1}p(x_{t+1}|y_{t+1})$$

ii. 计算过程

$$\alpha(y_1=1)=0.4*0.7=0.28$$
$$\alpha(y_1=2)=0.4*0.5=0.2$$
$$\alpha(y_1=3)=0.2*0.4=0.08$$

------------------------------------------------------------

$$\alpha(y_2=1)=\big((0.28*0.3)+(0.2*0.2)+(0.08*0.1)\big)*0.3$$
$$=(0.084+0.04+0.008)*0.3=0.0396$$
$$\alpha(y_2=2)=0.5*\big((0.28*0.5)+(0.2*0.3)+(0.08*0.4)\big)$$
$$=0.5*(0.14+0.06+0.032)=0.116$$
$$\alpha(y_2=3)=0.6*(0.28*0.2+0.2*0.5+0.08*0.5)$$
$$=0.6*(0.056+0.1+0.04)=0.1176$$

------------------------------------------------------------

$$\alpha(y_3=1)=0.7*(0.0396*0.3+0.116*0.2+0.1176*0.1)$$
$$=0.7*(0.01188+0.0232+0.01176)=0.032788$$
$$\alpha(y_3=2)=0.5*(0.0396*0.5+0.116*0.3+0.1176*0.4)$$
$$=0.5(0.0198+0.0348+0.04704)=0.05082$$
$$\alpha(y_3=3)=0.4*(0.0396*0.2+0.116*0.5+0.1176*0.5)$$
$$=0.4*(0.00792+0.058+0.0588)=0.04988$$

------------------------------------------------------------

$$\boldsymbol{p(x) = \sum_i \alpha(y_3 = i) = 0.032788 + 0.05082 + 0.04988 = 0.133488}$$

-------------------------------------------------------------

假设我们需要计算最佳状态序列

$$\beta(y_3) = 1\,?$$

-------------------------------------------------------------

$$\beta(y_2 = 1) = (0.3 * 0.7 + 0.5 * 0.5 + 0.2 * 0.4) = 0.21 + 0.25 + 0.08 = 0.54$$
$$\beta(y_2 = 2) = (0.2 * 0.7 + 0.3 * 0.5 + 0.5 * 0.4) = 0.14 + 0.15 + 0.2 = 0.49$$
$$\beta(y_2 = 3) = (0.1 * 0.7 + 0.4 * 0.5 + 0.5 * 0.4) = 0.07 + 0.2 + 0.2 = 0.47$$

-------------------------------------------------------------

$$\beta(y_1 = 1) = (0.54 * 0.3 * 0.3 + 0.49 * 0.5 * 0.5 + 0.47 * 0.2 * 0.6)$$
$$= 0.0486 + 0.1225 + 0.0564 = 0.2275$$
$$\beta(y_1 = 2) = (0.54 * 0.2 * 0.3 + 0.49 * 0.3 * 0.5 + 0.47 * 0.5 * 0.6)$$
$$= 0.0324 + 0.0735 + 0.141 = 0.2469$$
$$\beta(y_1 = 3) = (0.54 * 0.1 * 0.3 + 0.49 * 0.4 * 0.5 + 0.47 * 0.5 * 0.6)$$
$$= 0.0162 + 0.098 + 0.141 = 0.2552$$

-------------------------------------------------------------

$$p(y_3 = 1|x) = \frac{\alpha(y_3 = 1)}{p(x)} = \frac{0.030436}{0.11329} = 0.02686$$

$\cdots$ 显然$p(y_3 = 3|x)$ 最大

-------------------------------------------------------------

$$p(y_2 = 1|x) = \frac{\alpha(y_2 = 1)\beta(y_2 = 1)}{p(x)} = \frac{0.0396 * 0.54}{p(x)} = \frac{0.02138}{p(x)}$$
$$p(y_2 = 2|x) = \frac{\alpha(y_2 = 2)\beta(y_2 = 2)}{p(x)} = \frac{0.116 * 0.49}{p(x)} = \frac{0.05684}{p(x)}$$

$$p(y_2 = 3|x) = \frac{\alpha(y_2 = 3)\beta(y_2 = 3)}{p(x)} = \frac{0.1176 * 0.47}{p(x)} = \frac{0.055272}{p(x)}$$

显然$p(y_2 = 2|x)$ 最大

$p(x) = 0.02138 + 0.05684 + 0.055272 = 0.133492$

-------------------------------------------------------------

$$p(y_1 = 1|x) = \frac{\alpha(y_1 = 1)\beta(y_1 = 1)}{p(x)} = \frac{0.28 * 0.2275}{p(x)} = \frac{0.0637}{p(x)}$$
$$p(y_1 = 2|x) = \frac{\alpha(y_1 = 2)\beta(y_1 = 2)}{p(x)} = \frac{0.2 * 0.2469}{p(x)} = \frac{0.04938}{p(x)}$$
$$p(y_1 = 3|x) = \frac{\alpha(y_1 = 3)\beta(y_1 = 3)}{p(x)} = \frac{0.08 * 0.2552}{p(x)} = \frac{0.020416}{p(x)}$$

可见$p(y_1 = 1|x)$ 最大

$p(x) = 0.0637 + 0.04938 + 0.020416 = 0.133496$

最佳状态序列是　1→2→3，由于 3 个$p(x)$在千分位保持一致。