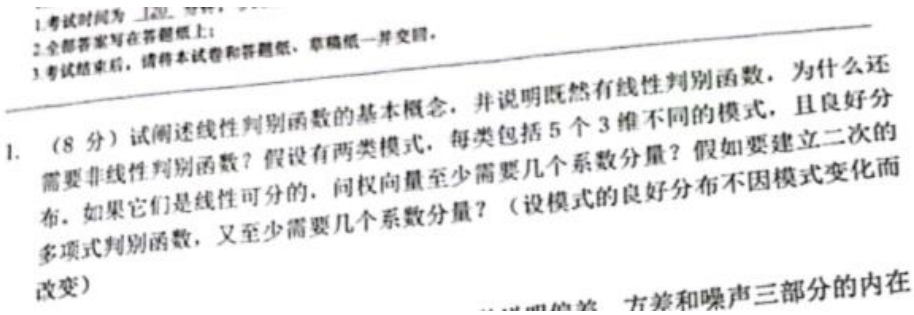

1. 考试时间为 [illegible]
2. 全部答案写在答题纸上；
3. 考试结束后，请将本试卷和答题纸、草稿纸一并交回。

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 5 个 3 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）

[illegible] 方差和噪声三部分的内在

i. 线性判别函数一般是$y = wx$　其中x是特征向量的增广形式，w 是权重系数。根据y的取值进行分类，这个函数在几何上一般表现为直线（高维空间的超平面），所以称之为线性判别函数。如果x是低纬向高维投影后的特征向量，那么就是广义线性判别，理论上广义线性判别可以模拟任意复杂的函数。

ii. 虽然广义线性判别可以拟合非线性，但是会面临参数爆炸问题，假如采用核技巧，虽然参数不会爆炸，但是面临 kernel 函数形式有限和无法度量哪个 kernel 函数更有效的问题，因此，假如经过先验知识，可以设计更合适的非线性的模型，是必要的。

iii. 参数数目

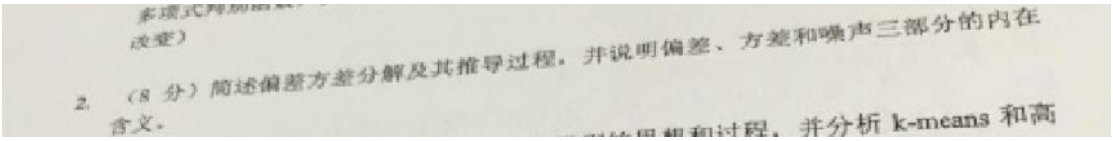
a) 线性需要（3+1）= 4

b) 二次需要（3(一次) + 3（二次） + 3(混合) + 1 （偏移）) = 10

c) 或者直接用：公式$\frac{(n+r)!}{n!r!}$

i. 线性 $n = 3, r = 1$ ； $\frac{(n+r)!}{n!r!} = 4$

ii. 二次$n = 3, r = 1$ ； $\frac{(n+r)!}{n!r!} = 10$


改变）

2. （8 分）简述偏差方差分解及其推导过程，并说明偏差、方差和噪声三部分的内在含义。

[illegible] 并分析 k-means 和高

i. 偏差-方差分解的推导过程

$$E(f_D, y_D) = E((f_D - y_D)^2) = E((f_D - f + f - y_D)^2)$$
$$= E((f_D - f)^2) + E((f - y_D)^2) + E(2(f_D - f)(f - y_D))$$

由于$E(f_D - f) = 0$,因此第三项为 0

$$= E((f_D - f)^2) + E((f - y + f - y_D)^2)$$
$$= E((f_D - f)^2) + E((f - y)^2) + E((y - y_D)^2) + E(2(f - y)(y - y_D))$$

由于$E(y - y_D) = 0$,因此第四项为 0

$$= E((f_D - f)^2) + E((f - y)^2) + E((y - y_D)^2)$$

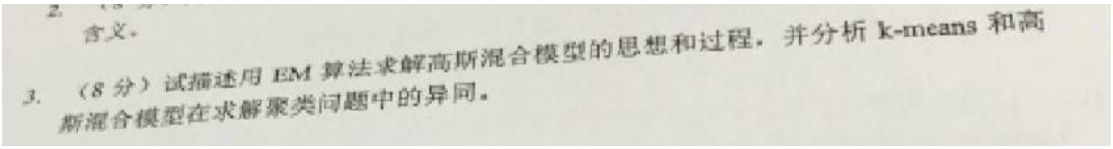
$E((f_D - f)^2)$ 是方差

$E((f - y)^2)$ 是偏差

$E((y - y_D)^2)$ 是误差

ii. 内在含义

a) 偏差：偏差是训练所使用的模型和模式之间的差异导致的错误
b) 方差：方差是相同模型在不同采样数据下训练带来的误差。
c) 噪音：是采样过程中采样误差（噪声）导致的训练结果的错误。


含义。

3. （8 分）试描述用 EM 算法求解高斯混合模型的思想和过程，并分析 k-means 和高斯混合模型在求解聚类问题中的异同。

i. EM 算法求解高斯混合模型
   a) 思想： 假定存在 M 个独立的高斯分布，数据x_i按照π_i 的概率从第i个高斯进行采样获取的。
   b) 过程
      i. 首先初始化混合高斯的参数$\theta = \{\pi\ , \delta, \mu\}$
      ii. 迭代直到收敛
         1. E 步骤，计算$\gamma(z_i^j) = p(y_i|x_i, \theta) = \frac{p(y_i|x_i,\theta)}{\sum_{i=1}^{M} p(y_i|x_i,\theta)}$
         2. M 步骤，更新

$$\theta = \ maxarg_\theta \log(p(x, y|\theta))$$
$$= maxarg_\theta \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} p(x_i, y_i^j|\ \theta) \right)$$

参数迭代公式（背过吧~）

$$\pi_k = \frac{\sum_i \gamma(z_i^k)}{N}$$
$$\mu_k = \frac{\sum_i \gamma(z_i^k) x_i}{\sum_i \gamma(z_i^k)}$$
$$\Sigma_k = \frac{\sum_i \gamma(z_i^k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i \gamma(z_i^k)}$$

ii. k-means 和高斯混合的异同
   a) 不同：
      i. k-means 的损失函数是最小平方距离， 混合高斯是负对数似然函数
      ii. k-means 是硬划分，混合高斯是软化分
      iii. k-maans 假设类别是概率相同且是球簇，混合高斯可以处理非球形，类别概率不同
   b) 相同：
      i. 混合高斯的 E 步骤其实就是软化分的 k-means
      ii. 当类别概率相同，且$\Sigma = \sigma I\ , \sigma \to 0\ , r_{i,j} \to \{0,1\}$ 的时候，混合高斯退化为 k-means。

---

4. （10分）用下列势函数

$$K(x,x_k)=e^{-\|x-x_k\|^2}$$

求解以下模式的分类问题

ω_1: $\{(0\ 1)^T,\ (0\ -1)^T\}$

ω_2: $\{(1\ 0)^T,\ (-1\ 0)^T\}$

迭代直到全部可以分类:

$$K_0(\boldsymbol{x})=0\quad;$$

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

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

$$K(x)=\exp(-(\boldsymbol{x}-(0,1)^T)^2)+\ \exp(-(\boldsymbol{x}-(0,-1)^T)^2)$$

$$-\exp(-(\boldsymbol{x}-(1,0)^T)^2)-\exp(-(\boldsymbol{x}-(-1,0)^T)^2)$$

刚好四个点都需要添加进去。

---

5. （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。

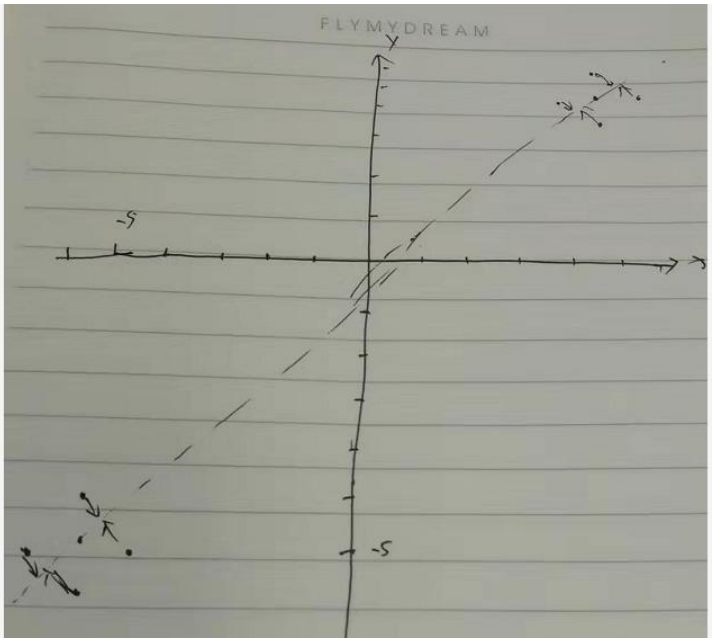
ω_1: $\{(-5\ -5)^T,\ (-5\ -4)^T,\ (-4\ -5)^T,\ (-5\ -6)^T,\ (-6\ -5)^T\}$

ω_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 的关注问题杀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$E[X]$
      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) $C=X^TX=\begin{Bmatrix}254 & -250\\ -250 & 254\end{Bmatrix}$
   c) $(C-\lambda I)x=0\ \rightarrow\begin{vmatrix}254-\lambda & -250\\ -250 & 254-\lambda\end{vmatrix}=0\rightarrow\begin{vmatrix}4-\lambda & 4-\lambda\\ 250 & 254-\lambda\end{vmatrix}=0$
      i. $\rightarrow\lambda=4\rightarrow$ 特征向量$(1,1)^T$
   d) 降维： $W=(1,1)^T$ ；
      i. $W^TX=\{\ \ -10\ ,-9\ ,-9\ ,-11\ ,-11,10,11,11,9\ ,9\ \}$

e) 

---

6. （10 分）详细描述 AdaBoost 算法，并解释为什么 AdaBoost 经常可以在训练误差为 0 后继续训练还可能带来测试误差的继续下降。

i. AdaBoost 算法

a) 预设样本为(x_i, y_i) ,$i = 1,2, \dots N$ ； 第 m 个弱分类器为$\phi_m(x)$ ，分类器性能为ε_m ， 分类器权重为α_m ，第 m 轮样本权重为$w_{m,i}$

b) 初始化样本权重$w_{1,i} = \frac{1}{N}$

c) 迭代 $m = 1{:}M$

a) 根据当前权重$w_{m,i}$ 训练弱分类器$\phi_m(x)$ ，要求性能优于随即猜想

b) 计算$\varepsilon_m = \sum_{i=1}^{N} w_{m,i} \ \mathbb{I}(y_i != \phi_m(x_i))$ ，$\alpha_m = \frac{1}{2}\log((1-\varepsilon_m)/\varepsilon_m)$

c) 更新$w_{m+1,i} = \frac{w_{m,i}\exp(-\alpha_m y_i \ \phi_m(x_i))}{Z_m}$ 其中Z_m是归一化因子

d) 最终的分类器是$sgn(\sum_m(\alpha_m\phi_m(x)))$

ii. 继续训练类似于增大 margin，虽然训练正确率 100%，但是泛化能力增加。

---

7. （10 分）描述感知机（Perceptron）模型，并给出其权值学习算法。在此基础上，以仅有一个隐含层的三层神经网络为例，形式化描述 Back-Propagation（BP）算法中是如何对隐层神经元与输出层神经元之间的连接权值进行调整的。

i. 感知器模型：

a) 描述： 感知器模型是一种赏罚模型，分类正确就不处罚，分类错误就处罚，

直到全部样本都分类正确为止。

b) 公式解释： $y = w^T x$ 期望对于所有x 有 $(y = w^T x) > 0$ ， 因此其损失函数为$J(w) = \sum_{y \in Y} y = \sum_{y_i \in Y} w^T x_i$ 其中Y 是全部错分$(y \leq 0)$ 样本。因此其权重更新公式为 $w_{i+1} = w_i + \eta \sum_{y_i \in Y} x_i$ ，其中η是超参数“步长”。

ii. BP 算法需要误差σ 的反向传播。预先定义：每一层的输入向量分别是 x ，y ，z ，真实标签是t ，每一层的输出是$f_{net}(y)$, $f_{net}(z)$ ，转移函数是f_{net} ， 权重是$w_{i\ ,h}$ 和$w_{\mathrm{h,g}}$

a) 在输出层，误差$\sigma_z = t - f_{net}(z)$ ，对应的输入的误差是$\sigma_{z_in} = \sigma_z f_{net}(z)'$

i. 隐含层→ 输出层的连接权值更新:

$$\Delta w_{\mathrm{h,g}} = {\sigma_{z\_in}}^g * \eta x_{\mathrm{h,g}} = (t^g - f_{net}(z)^g) f_{net}(z)^{g'} \eta x_{\mathrm{h,g}}$$

(如果题干要输入→ 隐含层的权重更新)

b) 在隐含层，输出的误差是输出层输入的加权求和 $\sigma_y = \sum_g w_{\mathrm{h,g}}\ {\sigma_{z\_in}}^g$

i. 对应的输入的误差是$\sigma_{y_in} = \sigma_y f_{net}(y)'$

ii. 对于输入层→ 隐含层的权重更新$\Delta w_{\mathrm{i,h}} = {\sigma_{y_in}}^g * \eta x_{\mathrm{i,h}}$

---

8. （12 分）已知正例点$\mathbf{x}_1 = (3,3)^{\mathrm{T}}, \mathbf{x}_2 = (4,3)^{\mathrm{T}}$，负例点$\mathbf{x}_3 = (1,1)^{\mathrm{T}}$，试用线性支持向量机的对偶算法求最大间隔分离超平面和分类决策函数，并在图中画出分离超平面、间隔边界及支持向量。

i. 对偶问题算法

$$\max \quad \sum_{i=1}^{n} \alpha^i - \frac{1}{2} \sum_{i,j=0}^{n} \alpha^i \alpha^j y^i y^j (x^i)^T x^j$$

$$st \quad \alpha^i \geq 0\ , i = 1\ , 2\ , \ldots n$$

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

显然 $\alpha^1 = 1$ ， $\alpha^2 = 0$ ，$\alpha^3 = 1$

$$w = \sum_{i=0}^{n} \alpha^i y^i x^i = (2,2)^T$$

支持向量是x^1, x^3 ,因此 $\frac{y^1}{||w||} = 1\ ; \rightarrow y^1 = 4$

$$b = y^1 - \sum_{j=0}^{n} \alpha^j y^j (x^i)^T x^j = 4 - \quad 18 + \quad 6 = -8$$

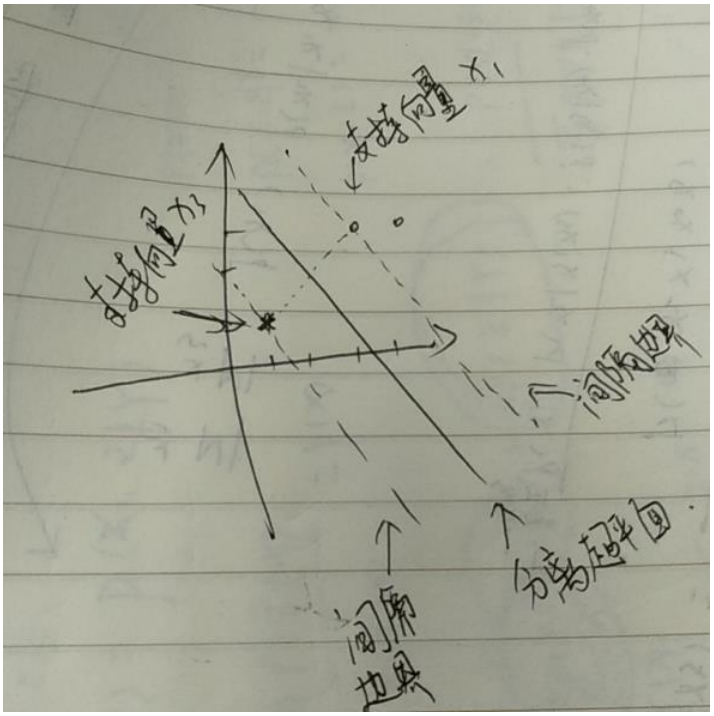
同理$y^3 = -4$

$b = y^3 - \sum_{j=0}^{n} \alpha^j y^j (x^i)^T x^j = -4 - 6 + 2 = -8$

无论从哪个点计算都是一样的，因此：

$(2,2)^T x - 8 = 0$ ← 分类面。二维坐标系下 ：$x + y - 4 = 0$

ii. 草图



间隔边界及支持向量。

9. （12 分）假定对一类特定人群进行某种疾病检查，正常人以 ω_1 类代表，患病者以 ω_2 类代表。设被检查的人中正常者和患病者的先验概率分别为

正常人：$P(\omega_1)=0.9$

患病者：$P(\omega_2)=0.1$

现有一被检查者，其观察值为 x，从类条件概率密度分布曲线上查得

$P(x|\omega_1)=0.2$，$P(x|\omega_2)=0.4$

同时已知风险损失函数为

$$\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix}=\begin{pmatrix}0 & 6\\ 1 & 0\end{pmatrix}$$

其中 λ_{ij} 表示将本应属于第 j 类的模式判为属于第 i 类所带来的风险损失。试对该被检查者用以下两种方法进行分类：

（1） 基于最小错误率的贝叶斯决策，并写出其判别函数和决策面方程；

（2） 基于最小风险的贝叶斯决策，并写出其判别函数和决策面方程。

i. 最小错误率贝叶斯：

a) 判别函数 ：

$$if\ \ p(x|w_1)*p(w_1) > \ \ p(x|w_2)*p(w_2) \rightarrow \ \ w_1\ ;$$
$$else\ \ if\ \ p(x|w_1)*p(w_1) < \ \ \ p(x|w_2)*p(w_2) \rightarrow \ \ w_2$$

b) 决策面方程

$$p(x|w_1)*p(w_1) - \ \ p(x|w_2)*p(w_2) = \ \ 0$$

c) 决策

$$p(x|w_1)*p(w_1) = 0.18$$
$$p(x|w_2)*p(w_2) = 0.04$$

判决属于w_1。

ii. 最小风险贝叶斯：

a) 判别函数

$$if\ \ p(x|w_2)*p(w_2)\lambda_{22}+p(x|w_1)*p(w_1)\ \ \lambda_{21}$$
$$<\ \ p(x|w_1)*p(w_1)\ \ \lambda_{11}+p(x|w_2)*p(w_2)\lambda_{12}\ \ \rightarrow w_2$$
$$if\ \ p(x|w_2)*p(w_2)\lambda_{22}+p(x|w_1)*p(w_1)\ \ \lambda_{21}$$
$$>\ \ p(x|w_1)*p(w_1)\ \ \lambda_{11}+p(x|w_2)*p(w_2)\lambda_{12}\ \ \rightarrow w_1$$

b) 决策面方程

$$p(x|w_2)*p(w_2)\ (\ \lambda_{22}-\lambda_{12})+p(x|w_1)*p(w_1)\ (\ \lambda_{11}-\lambda_{21})=\ \ 0$$

c) 决策：

$$p(x|w_2)*p(w_2)\lambda_{22}+p(x|w_1)*p(w_1)\ \ \lambda_{21}=0.18$$
$$p(x|w_1)*p(w_1)\ \ \lambda_{11}+p(x|w_2)*p(w_2)\lambda_{12}=0.24$$

故判决属于w_2

---

(2) 基于最小风险[illegible]

10. （12 分）假设有 3 个盒子，每个盒子里都装有红、白两种颜色的球。按照下面的方法抽球，产生一个球的颜色的观测序列：开始，以概率 π 随机选取 1 个盒子，从这个盒子里以概率 B 随机抽出 1 个球，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；如此重复进行 3 次，得到一个球的颜色观测序列：O =（红，白，红）。请计算生成该序列的概率P(O|{A, B, π})。

提示：假设状态集合是{盒子 1，盒子 2，盒子 3}，观测的集合是{红，白}，本题中已知状态转移概率分布、观测概率分布和初始概率分布分别为：

$$A=\begin{matrix}\\ 盒子1\\ 盒子2\\ 盒子3\end{matrix}\begin{matrix}盒子1\ \ 盒子2\ \ 盒子3\\ \begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}\end{matrix},\quad B=\begin{matrix}\\ 盒子1\\ 盒子2\\ 盒子3\end{matrix}\begin{matrix}红\ \ 白\\ \begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}\end{matrix},\ \pi=[0{,}2, 0.4, 0.4]^T。$$

i. 前向计算需要的公式

$$\alpha(t+1)=\sum_{y_t}\alpha(t)\ \ A_{y_t,y_t+1}\ \ B_{y_t+1,x}$$
$$=B_{y_t+1,x}\sum_{y_t}\alpha(t)\ \ A_{y_t,y_t+1}$$

ii. 计算

$$\alpha(y_t=1\ \ ,t=1)=\pi_1*p(x=红|y=1)=0.1$$

$$\alpha(y_t=2\ \ ,t=1)=0.4*0.4=0.16$$
$$\alpha(y_t=3,t=1)=0.4*0.7=0.28$$

$$\alpha(y_t=1\ \ ,t=2)=0.5*\ \ (0.1*0.5+0.16*0.3+0.28*0.2)$$
$$=0.5*(0.05+0.048+0.056\ \ )=0.077$$

$$\alpha(y_t = 2\ , t = 2) = 0.6 * (0.1 * 0.2 + 0.16 * 0.5 + 0.28 * 0.3)$$
$$= 0.1104$$
$$\alpha(y_t = 3\ , t = 2) = 0.3 * (0.1 * 0.3 + 0.16 * 0.2 + 0.28 * 0.5) = 0.0606$$

$$\alpha(y_t = 1\ , t = 3) = 0.5 * \ (0.077 * 0.5 + 0.1104 * 0.3 + 0.0606 * 0.2)$$
$$= 0.5 * (0.0385 + 0.03312 + 0.01212) = 0.04187$$
$$\alpha(y_t = 2\ , t = 2) = 0.4 * (0.077 * 0.2 + 0.1104 * 0.5 + 0.0606 * 0.3)$$
$$= 0.4 * (0.0154 + 0.0552 + 0.01818) = 0.035512$$
$$\alpha(y_t = 3\ , t = 3) = 0.7 * (0.077 * 0.3 + 0.1104 * 0.2 + 0.0606 * 0.5)$$
$$= 0.7 * (\ 0.0231 + 0.02208 + 0.0303\ ) = 0.05283$$

$$p(x) = \sum_{y_t} \alpha(y_t\ , t = 3)\ = 0.13021$$

因为没有仔细检查，上述计算可能存在错误！关键是知道前向计算的迭代公式！

1. （6 分）简述模式的概念和它的直观特性，并简要说明模式分类有哪几种主要方法。

i. 广义的说，存在于时间和空间中可观测的物体，如果我们可以区别它们是否相同或者相似，都可以称之为模式。模式所指的不是事物本身，而是从事物获得的信息。因此模式往往指的是具有时间或空间分布的信息。

ii. 模式的直观特征：可观察性，可区分性，相似性

iii. 主要方法：

a) 监督学习：概念驱动，归纳假说。

b) 非监督学习： 数据驱动，演绎假说。

---

2. （8分）假设某研究者在ImageNet数据上使用线性支持向量机（Linear SVM）来做文本分类的任务，请说明在如下情况下分别如何操作才能得到更好的结果，并说明原因。

（1）训练误差5%，验证误差10%，测试误差10%。

（2）训练误差1%，验证误差10%，测试误差10%。

（3）训练误差1%，验证误差3%，测试误差10%。

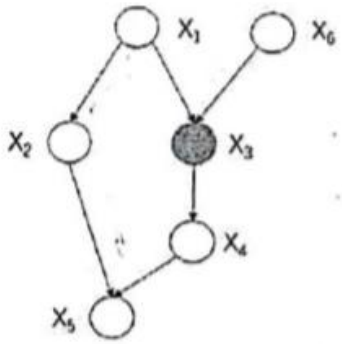
i. 欠拟合，适当的增大 C 值，减少错分样本。

ii. 过拟合，适当的降低 C 值，增加模型的泛化能力。

iii. 训练数据和测试数据不是独立同分布，建议重新采样或者 shuffle 数据。

---

3. （8分）给定如下概率图模型，其中变量X_3为已观测变量，请问变量X_4和X_6是否独立？并用概率推导证明之。



$$p(x1,x2,x3,x4,x5,x6) = p(x1) * p(x6) * p(x3|x1,x6) * p(x2|x1) * p(x4|x3) * p(x5|x4,x2)$$

$$p(x3,x4,x6) = \sum_{x1}\sum_{x2}\sum_{x5} p(x1) * p(x6) * p(x3|x1,x6) * p(x2|x1) * p(x4|x3) * p(x5|x4,x2)$$

$$= \sum_{x1} p(x1) * p(x6) * p(x3|x1,x6) * p(x4|x3) * \sum_{x2} p(x2|x1) * \sum_{x5} p(x5|x4,x2)$$

$$= \sum_{x1} p(x1) * p(x6) * p(x3|x1,x6) * p(x4|x3)$$

$$p(x3,x6)=\sum_{x1}\sum_{x4}p(x1)*p(x6)*p(x3|x1,x6)*p(x4|x3)$$
$$=\sum_{x1}p(x1)*p(x6)*p(x3|x1,x6)\sum_{x4}p(x4|x3)$$
$$=\sum_{x1}p(x1)*p(x6)*p(x3|x1,x6)$$
$$p(x4|x3,x6)=\frac{p(x4,x3,x6)}{p(x3,x6)}$$
$$=\frac{\sum_{x1}p(x1)*p(x6)*p(x3|x1,x6)*p(x4|x3)}{\sum_{x1}p(x1)*p(x6)*p(x3|x1,x6)}$$
$$=\ p(x4|x3)$$

得证$x3$已知的情况下，$x4$和$x6$独立。

---

4. （10 分）（1）随机猜测作为一个分类算法是否一定比 SVM 差？借此阐述你对“No Free Lunch Theorem”的理解。（2）举例阐述你对“Occam's razor”的理解。

i. 脱离具体问题谈论算法优劣是没有意义的，在特定的问题上随机猜想是可以比 SVM 好的。

ii. No Free Lunch Theorem：在问题等概率出现且等权重的情况下，任何算法的期望都是一样的。也就是说，没有一个算法可以在任何问题上总是产生最好的分类器。脱离具体问题讨论算法的优劣是无意义的。只有针对具体问题的具体模型，才能对比优劣。

iii. Occam's razor ：这是一种归纳偏好： 如无必要，勿增实体。达到相近性能的模型中，最简单的往往更加接近真相。过度复杂只会造成过拟合而失去泛化能力。

---

5. （10 分）详细描述 AdaBoost 的原理并给出算法，并解释为什么 AdaBoost 经常可以在训练误差为 0 后继续训练还可能带来测试误差的继续下降。

i. AdaBoost 原理： 基于强分类器比较难以获取，期望训练多个弱分类器配合构成一个强分类器的思想，AdaBoost 使用在弱分类器 1 上训练失败的样本去训练弱分类器 2 的思路，通过调整样本权重，使得弱分类器 1 在样本上等价于随即猜想，然后用调整后的权重样本去训练分类器 2.

ii. AdaBoost 算法：

a) 初始化样本权重$w_{1,i}=\frac{1}{N}$

b) 迭代 $m=1:M$

i. 在$w_{m,i}$权重下训练弱分类器$\emptyset_m(x)$ ，要求分类器性能优于随即猜想。

ii. 计算$\varepsilon=\sum_i w_{m,i}\mathbb{I}(\emptyset_m(x_i)\ !=y_i\ )$

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

1. 其中$\alpha_m = \frac{1}{2}(log\frac{(1-\varepsilon)}{\varepsilon})$ ；Z_m是归一化因子。

c) 最终的训练器是$sgn\left(\sum_m \alpha_m \emptyset_m(x)\right)$

iii. 训练误差为 0 后 AdaBoost 继续训练类似于继续寻找更大的分类 margin

---

6. （10 分）用感知器算法求下列模式分类的解向量（取 w(1)为零向量）

ω_1: $\{(0\ 0\ 0)^T,\ (1\ 0\ 0)^T,\ (1\ 0\ 1)^T,\ (1\ 1\ 0)^T\}$

ω_2: $\{(0\ 0\ 1)^T,\ (0\ 1\ 1)^T,\ (0\ 1\ 0)^T,\ (1\ 1\ 1)^T\}$

i. 获得规范增广矩阵

$(0,0,0,1)^T$ ；$(\ 1,0,0,1\ )^T$ ；$(\ 1,0,1,1)^T$ ；$(\ 1,1,0,1\ )^T$

$(0,0,-1,-1)^T$；$(\ 0,-1,-1,-1)^T$ ；$(\ 0,-1,0,-1)^T$；$(\ -1,-1,-1,-1)^T$

ii. 初始化$w = (0,0,0,0)^T$

iii. 迭代

a) 第一轮全军覆没 $w = (2,-2,-2,0)^T$

b) 第二轮 $(0,0,0,1)^T$ $(\ 1,0,1,1)^T$ $(\ 1,1,0,1\ )^T$ 错误 $w = (2,-1,-1,3)^T$

c) 第三轮第二列错误 $w = (1,-4,-4,-1)^T$

d) 第四轮第一列错误$w = (4,-3,-3,3)^T$

e) 第五轮$(0,0,-1,-1)^T$ $(\ 0,-1,0,-1)^T$ $(\ -1,-1,-1,-1)^T$ 错误，$w = (3,-5,-5,0)^T$

f) 第六轮$(0,0,0,1)^T(\ 1,0,1,1)^T$ ；$(\ 1,1,0,1\ )^T$ 错误$w = (3,-4,-4,3)^T$

g) 第七轮全部完成，解向量$w = (3,-4,-4,3)^T$

---

7. （12 分）设以下模式类别具有正态概率密度函数：

ω_1: $\{(0\ 0\ 0)^T,\ (1\ 0\ 0)^T,\ (1\ 0\ 1)^T,\ (1\ 1\ 0)^T\}$

ω_2: $\{(0\ 1\ 0)^T,\ (0\ 1\ 1)^T,\ (0\ 0\ 1)^T,\ (1\ 1\ 1)^T\}$

若 $P(\omega_1)=P(\omega_2)=0.5$，求这两类模式之间的贝叶斯判别界面的方程式。

$$u_1 = \frac{1}{4}(\ 3,1,1)$$

$$\Sigma_1 = (w_1 - u_1)^T(w_1 - u_1)\ ;$$

$$(w_1 - u_1)^T = \begin{Bmatrix} -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \end{Bmatrix}$$

$$\Sigma_1 = \begin{Bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{Bmatrix}$$

$$u_2 = \frac{1}{4}(\ 1,3,3)$$

$$\Sigma_2 = (w_2 - u_2)^T (w_2 - u_2)$$

$$(w_2 - u_2)^T = \begin{Bmatrix} -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} \\ -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix}$$

$$\Sigma_2 = \begin{Bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{Bmatrix}$$

可见，$\Sigma_1 = \Sigma_2$ 又由于$p(w_1) = p(w_2)$，因此这是一个最小马氏距离分类器。

分类界面方程 $(x - u_1)^T \Sigma_1 (x - u_1) = \ (x - u_2)^T \Sigma_2 (x - u_2)$

$$y = wx + x_0$$

$$x_0 = \frac{1}{2}(u_2 + u_1) = (0.5,0.5,0.5)$$

$$w = \Sigma^{\ -1}(u_1 - u_2\ )$$

就不继续计算了！give up

---

8. （12 分）假设有如下线性回归问题，

$$\min_{\beta}(y - X\beta)^2 + \lambda||\beta||_2^2$$

其中y和β是n维向量，X是一个$m \times n$的矩阵。

该线性回归问题的参数估计可看作一个后验分布的均值，其先验为高斯分布$\beta \sim N(0, \tau I)$，样本产生自高斯分布$y \sim N(X\beta, \sigma^2 I)$，其中$I$为单位矩阵，试推导调控系数$\lambda$与方差$\tau$和$\sigma^2$的关系。

$$p(\beta|y^{\rightarrow}) = \frac{p(\beta, y^{\rightarrow})}{p(y^{\rightarrow})} = \frac{p(\beta|\ \tau)\ p(y^{\rightarrow}|\beta, X, \sigma)}{p(y^{\rightarrow})}$$

$$\log\big(p(\beta|y^{\rightarrow})\big) = \log(\ p(\beta|\ \tau)) + \log\big(\ p(y^{\rightarrow}|\beta, X, \sigma)\big)\ - \log\big(p(y^{\rightarrow})\big)$$

$$= \log\left(\ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi||\tau I||}} exp\left(-\frac{1}{2}\beta^T(\tau I)^{-1}\beta\right)\ \right)$$

$$+ \log\left(\ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi||\sigma I||}} exp\left(-\frac{1}{2}\big(y^i - x\beta\big)^T(\sigma I)^{-1}\big(y^i - x\beta\big)\right)\ \right) - \log\big(p(y^{\rightarrow})\big)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi||\tau I||}} exp\left(-\frac{1}{2}\beta^T(\tau I)^{-1}\beta \right) \right)$$

$$+ \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi||\sigma I||}} exp\left(-\frac{1}{2}\left(y^i - x\beta\right)^T(\sigma I)^{-1}\left(y^i - x\beta\right) \right) \right) - \log\left(p(y^{\rightarrow})\right)$$

$$= Nlog\left(\frac{1}{\sqrt{2\pi||\sigma I||}} \right) + \sum_{i=1}^{n} \log\left(exp\left(-\frac{1}{2}\beta^T(\tau I)^{-1}\beta \right) \right) + Nlog\left(\frac{1}{\sqrt{2\pi||\sigma I||}} \right)$$

$$+ \sum_{i=1}^{n} \log\left(exp\left(-\frac{1}{2}\left(y^i - x\beta\right)^T(\sigma I)^{-1}\left(y^i - x\beta\right) \right) \right) - \log\left(p(y^{\rightarrow})\right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2}\left(y^i - x\beta\right)^T(\sigma I)^{-1}\left(y^i - x\beta\right) - \frac{1}{2}\beta^T(\tau I)^{-1}\beta \right) + const$$

$$= -\frac{1}{2}\sum_{i=1}^{n} \left(\left(y^i - x\beta\right)^T(\sigma I)^{-1}\left(y^i - x\beta\right) + \frac{1}{2}\beta^T(\tau I)^{-1}\beta \right) + C$$

$$\propto -\sum_{i=1}^{n} \left(\left(y^i - x\beta\right)^2 + \frac{\sigma}{\tau}||\beta||^2 \right) + C$$

因此，最大似然等价于min $\left(y - x\beta\right)^2 + \frac{\sigma}{\tau}||\beta||^2$,因此$\lambda = \frac{\sigma}{\tau}$

---

9. （12 分）给定有标记样本集$D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$和未标记样本$D_u = \{(x_{l+1}, y_{l+1}), (x_{l+2}, y_{l+2}), \ldots, (x_{l+u}, y_{l+u})\}, l \ll u, l + u = m$，假设所有样本独立同分布，且都是由同一个包含 N 个混合成分的高斯混合模型$\{(\alpha_i, \mu_i, \Sigma_i)|1 \leq i \leq N\}$产生，每个高斯混合成分对应一个类别，请写出极大似然估计的目标函数（对数似然函数），以及用 EM 算法求解参数的迭代更新式。

i. 极大似然估计的目标函数

令$\theta = (\alpha, \mu, \Sigma)$

$$\log\ p(X_L, X_u, Y_L \mid \theta) = \log\left(\prod_{i=0}^{l} p(x_i, y_i|\ \theta) * \prod_{i=l+1}^{l+u} \sum_{j=1}^{N} p(x_i, y_j|\ \theta) \right)$$

$$= \sum_{i=0}^{l} \log(p(\ x_i|y_i\ \theta)p(y_i|\ \theta)) + \sum_{i=l+1}^{l+u} \log(\ \sum_{j=1}^{N} p((x_i|y_j\ \theta)p(y_j|\ \theta)$$

ii. EM 算法求参数的迭代方式

a) 初始化一个θ

b) 迭代

i. 根据当前θ求y_j的分布（无标签部分）

1. $\gamma(z_i^j) = p(y_j|x_i,\theta) = \frac{p((x_i|y_j\ \theta)p(y_j|\ \theta)}{\sum_{j=1}^{N} p((x_i|y_j\ \theta)p(y_j|\ \theta)}$

ii. 有标签部分$\gamma(z_i^j) = 1\ if\quad y_j = 1\ ; else\ \gamma(z_i^j) = 0$

iii. 利用θ的最大似然估计，用估计值更新θ

参数迭代公式（背过吧~）

$$\pi_k = \frac{\sum_i \gamma(z_i^k)}{N}$$

$$\mu_k = \frac{\sum_i \gamma(z_i^k)x_i}{\sum_i \gamma(z_i^k)}$$

$$\Sigma_k = \frac{\sum_i \gamma(z_i^k)(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_i \gamma(z_i^k)}$$

---

10. （12 分）假定对一类特定人群进行某种疾病检查，正常人以 ω_1类代表，患病者以 ω_2类代表。设被检查的人中正常者和患病者的先验概率分别为

正常人：$P(\omega_1)=0.9$

患病者：$P(\omega_2)=0.1$

现有一被检查者，其观察值为 x，从类条件概率密度分布曲线上查得

$P(x|\omega_1)=0.2$，$P(x|\omega_2)=0.4$

同时已知风险损失函数为

$$\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix}=\begin{pmatrix}0 & 6\\ 1 & 0\end{pmatrix}$$

其中 λ_{ij}表示将本应属于第 j 类的模式判为属于第 i 类所带来的风险损失。试对该被检查者用以下两种方法进行分类：

（1） 基于最小错误率的贝叶斯决策，并写出其判别函数和决策面方程；

（2） 基于最小风险的贝叶斯决策，并写出其判别函数和决策面方程。

i. 最小错误率贝叶斯：

a) 判别函数 ：

$$if\ \ p(x|w_1)*p(w_1) > \ p(x|w_2)*p(w_2) \rightarrow \ w_1\ ;$$

$$else\ \ if\ \ p(x|w_1)*p(w_1) < \ \ p(x|w_2)*p(w_2) \rightarrow :\ w_2$$

b) 决策面方程

$$p(x|w_1)*p(w_1) - \ p(x|w_2)*p(w_2) = \ 0$$

c) 决策

$$p(x|w_1)*p(w_1) = 0.18$$

$$p(x|w_2)*p(w_2) = 0.04$$

判决属于w_1。

ii. 最小风险贝叶斯：

a) 判别函数

$$if\ \ p(x|w_2)*p(w_2)\lambda_{22} + p(x|w_1)*p(w_1)\ \ \lambda_{21}$$

$$< \ \ p(x|w_1)*p(w_1)\ \ \lambda_{11} + p(x|w_2)*p(w_2)\lambda_{12} \ \ \rightarrow w_2$$

$$if\ \ p(x|w_2) * p(w_2)\lambda_{22} + p(x|w_1) * p(w_1)\ \ \lambda_{21}$$
$$>\ \ p(x|w_1) * p(w_1)\ \ \lambda_{11} + p(x|w_2) * p(w_2)\lambda_{12}\ \ \rightarrow w_1$$

b) 决策面方程

$$p(x|w_2) * p(w_2)\ (\ \lambda_{22} - \lambda_{12}) + p(x|w_1) * p(w_1)\ (\ \lambda_{11} - \lambda_{21}) =\ \ 0$$

c) 决策：

$$p(x|w_2) * p(w_2)\lambda_{22} + p(x|w_1) * p(w_1)\ \ \lambda_{21} = 0.18$$
$$p(x|w_1) * p(w_1)\ \ \lambda_{11} + p(x|w_2) * p(w_2)\lambda_{12} = 0.24$$

故判决属于w_2

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 （$4(\text{一次项}) + 4(\text{二次项}) + C_4^2(\text{混合项}) + 1(w_0)$ ）

iii. 公式$\frac{(n+r)!}{n!r!}$

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

2. 二次$n = 4, r = 1$ ； $\frac{(n+r)!}{n!r!} = 15$

---

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：

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

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

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

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。

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

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

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

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。

ω_1: $\{(0\ 0)^T,\ (0\ 1)^T\}$

ω_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，时画出样本在该空间中的位置。

ω_1: $\{(-5\ -5)^T,\ (-5\ -4)^T,\ (-4\ -5)^T,\ (-5\ -6)^T,\ (-6\ -5)^T\}$

ω_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 的关注问题杀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$E[X]$

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

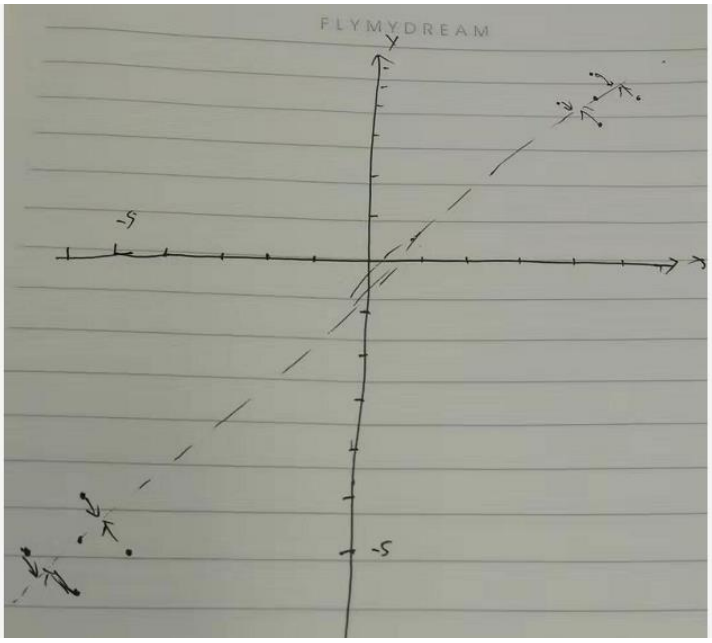
b) $C = XX^T = \begin{Bmatrix} 254 & -250 \\ -250 & 254 \end{Bmatrix}$

c) $(C - \lambda I)x = 0 \rightarrow \begin{vmatrix} 254-\lambda & -250 \\ -250 & 254-\lambda \end{vmatrix} = 0 \rightarrow \begin{vmatrix} 4-\lambda & 4-\lambda \\ 250 & 254-\lambda \end{vmatrix} = 0$

i. $\rightarrow \lambda = 4 \rightarrow$ 特征向量$(1,1)^T$

d) 降维： $W = (1,1)^T$；

i. $W^T X = \{-10, -9, -9, -11, -11, 10, 11, 11, 9, 9\}$

e) 

（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$w_{m+1,i} = \frac{w_{m,1}\exp(-\alpha_m y_i \emptyset_m(x_i))}{Z_m}$ 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特函数算法求解以下模式的分类问题

ω_1：$[(0\ 1)^T,\ (0\ -1)^T]$

ω_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:

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

---

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

$$p(y=1) = \frac{3}{8}$$
$$p(A=0) = 0.25$$

$$p(B=1)=0.25$$
$$p(A=0|Y=1)=0$$
$$p(A=1|Y=1)=1$$
$$p(A=0|Y=0)=0.5$$
$$p(A=0|Y=1)=0.5$$
$$p(B=0|Y=1)=0.75$$
$$p(B=1|Y=1)=0.25$$
$$p(B=0|Y=0)=0.8$$
$$p(B=1|Y=0)=0.2$$

$$p(Y=1 \mid A=0,B=1)=\frac{0*\frac{3}{8}*0.25*\frac{3}{8}}{0.25*0.25}=0$$

---

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 π 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 $P(O|\{A,B,\pi\})$。

$\pi=[0.4,0.4,0.2]^T$

$$A=\begin{array}{c} \\ \text{罐子1} \\ \text{罐子2} \\ \text{罐子3} \end{array}\begin{array}{c} \begin{array}{ccc}\text{罐子1} & \text{罐子2} & \text{罐子3}\end{array} \\ \begin{bmatrix}0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\\ 0.1 & 0.4 & 0.5\end{bmatrix}\end{array}$$

$$B=\begin{array}{c} \\ \text{罐子1} \\ \text{罐子2} \\ \text{罐子3} \end{array}\begin{array}{c} \begin{array}{cc}\text{红} & \text{黑}\end{array} \\ \begin{bmatrix}0.7 & 0.3\\ 0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix}\end{array}$$

i. 参考公式

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

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

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

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

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

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

… 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。

i. ***（10 分）简述 Fisher 线性判别方法的基本思路，写出准则函数和对应的解。***

1) Fisher 线性判别方法的基本思路 ：为了将d维空间内的样本投影到 1 维空间并且尽量保留可分性，Fisher 线性判别方法的基本思路是：选择最佳投影方向w^*，使得投影后各类样本内部尽量密集（也就是“类内散度”小），各类均值之差越大越好（也就是“类间散度“大）。

2) 准则函数（2 类问题）：

$$J(F)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2}=\frac{w^TS_bw}{w^TS_ww}$$

其中，

- $(\tilde{m}_1-\tilde{m}_2)$是**投影后**的两类均值之差；
- $\tilde{S}_i^2$是**投影后**的样本类内离散度；
  - $\tilde{S}_i^2=\sum_{y\in\Gamma_i}(y-\tilde{m}_i)^2$ 这是个标量，因为y是一维标量。
- w是投影方向；
- S_b为样本类间离散度矩阵；
  - $S_b=(m_1-m_2)(m_1-m_2)^T$
- S_w为总样本类内离散度矩阵；
  - $S_i=\sum_{x\in\Gamma_i}(x-m_i)(x-m_i)^T$ ，$i=1$ ，2
  - $S_w=S_1+S_2$

3) 解（2 类问题）：

$$S_w^{-1}S_bw^*=\lambda w^*$$

也就是说，w^*可以通过对$S_w^{-1}S_b$矩阵进行特征值分解获得，特殊的，在映射到 1 维的情况下： $w^*=S_w^{-1}(m_1-m_2)$。

---

ii. 2. （12 分）假设某个地区细胞识别中正常（w_1）和异常（w_2）两类的先验概率分别为：正常状态：$P(w_1)=0.95$，异常状态 $P(w_2)=0.05$。现有一待识别的细胞，其观察值为 x，已知 $p(x|w_1)=0.2$，$p(x|w_2)=0.5$。同时已知风险损失函数为：$\begin{pmatrix}\lambda_{11}&\lambda_{12}\\\lambda_{22}&\lambda_{21}\end{pmatrix}=\begin{pmatrix}0&1\\8&0\end{pmatrix}$

其中 λ_{ij} 表示将本应属于第 j 类的模式判为属于第 i 类所带来的风险损失。试对该待识别细胞用以下两种方法进行分类：

1) 基于最小错误率的贝叶斯决策，并写出其判别函数和决策面方程。
2) 基于最小风险的贝叶斯决策，并写出其判别函数和决策面方程。

1) 题干的损失矩阵可能是$\lambda_{12}=8$ $\lambda_{21}=1$；

最小错误率的贝叶斯决策：

i. 决策函数:

若 $p(w_1)p(x|w_1)>p(w_2)p(x|w_2)$ 则 w_1 。

若 $p(w_1)p(x|w_1)<p(w_2)p(x|w_2)$ 则 w_2 。

ii. 决策面方程

$$p(w_1)p(x|w_1)-p(w_2)p(x|w_2)=0$$

iii. 当前决策如下：

$p(w_1)p(x|w_1)=0.95*0.2=0.19>p(w_2)p(x|w_2)=0.05*0.5=0.025$ 故选w_1。

最小风险贝叶斯决策：

i. 决策函数：

若$p(w_1)p(x|w_1)\lambda_{11} + p(w_2)p(x|w_2)\lambda_{12} < p(w_1)p(x|w_1)\lambda_{21} + p(w_2)p(x|w_2)\lambda_{22}$，则$w_1$。
若$p(w_1)p(x|w_1)\lambda_{11} + p(w_2)p(x|w_2)\lambda_{12} > p(w_1)p(x|w_1)\lambda_{21} + p(w_2)p(x|w_2)\lambda_{22}$，则$w_2$。

ii. 判别界面

$$p(w_1)p(x|w_1)\lambda_{11} + p(w_2)p(x|w_2)\lambda_{12} - p(w_1)p(x|w_1)\lambda_{21} - p(w_2)p(x|w_2)\lambda_{22} = 0$$

$$p(w_1)p(x|w_1)(\lambda_{11} - \lambda_{21}) = p(w_2)p(x|w_2)(\lambda_{22} - \lambda_{12})$$

iii. 此案例判别

$$p(w_1)p(x|w_1)\lambda_{11} + p(w_2)p(x|w_2)\lambda_{12} = 0.2$$

$$p(w_1)p(x|w_1)\lambda_{21} + p(w_2)p(x|w_2)\lambda_{22} = 0.19$$

故选w_2。

---

3. （10 分）SVM 可以借助核函数（kernel function）在特征空间（feature space）学习一个具有最大间隔的超平面。对于两类的分类问题，任意输入x的分类结果取决于下式：

$$< \hat{w}, \phi(x) > + \hat{w}_0 = f(x; \alpha, \hat{w}_0)$$

其中，$\hat{w}$和$\hat{w}_0$是分类超平面的参数，$\alpha = [\alpha_1, \ldots \alpha_{|SV|}]$表示支持向量（support vector）的系数，SV表示支持向量集合。使用径向基函数（radial basis function）定义核函数$K(\cdot,\cdot)$，即$K(x, x') = \exp(-\frac{D(x,x')}{2s^2})$。假设训练数据在特征空间线性可分，SVM 可以完全正确地划分这些训练数据。给定一个测试样本x_{far}，它距离所有训练样本都非常远。

试写出$f(x; \alpha, \hat{w}_0)$在核特征空间的表达形式，进而证明：$f(x_{far}; \alpha, \hat{w}_0) \approx \hat{w}_0$

$$f(x; \alpha, \hat{w}_0) = w^* \Phi(x) + \hat{w}_0$$

$$= (\sum_{i=1}^{n} \alpha_i y_i x_i)^T * \Phi(x) + \hat{w}_0$$

$$= \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + \hat{w}_0$$

$$= \sum_{i=1}^{n} \alpha_i y_i \exp\left(-\frac{D(x_i - x)}{2s}\right) + \hat{w}_0$$

证明：

$$f(x_{far}; \alpha, \hat{w}_0) = \sum_{i=1}^{n} \alpha_i y_i \exp\left(-\frac{D(x_i - x_{far})}{2s}\right) + \hat{w}_0$$

由于$D(x_i - x_{far})$很大，所以$\exp\left(-\frac{D(x_i - x_{far})}{2s}\right)$趋近于 0，所以$f(x_{far}; \alpha, \hat{w}_0) \approx 0 + \hat{w}_0 = \hat{w}_0$

---

4. （10 分）K-L 变换属于有监督学习（supervised learning）还是无监督学习（unsupervised learning）？试利用 K-L 变换将以下样本集的特征维数降到一维，同时画出样本在该空间的位置。

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

i. K-L 变换属于无监督学习。

ii. 计算细节：

a) 首先计算样本均值 ： $m=(0,0)^T$ 符合最佳 K-L 变换需求

b) 然后计算$(x-m)(x-m)^T=\begin{pmatrix}254 & -250\\ -250 & 254\end{pmatrix}$

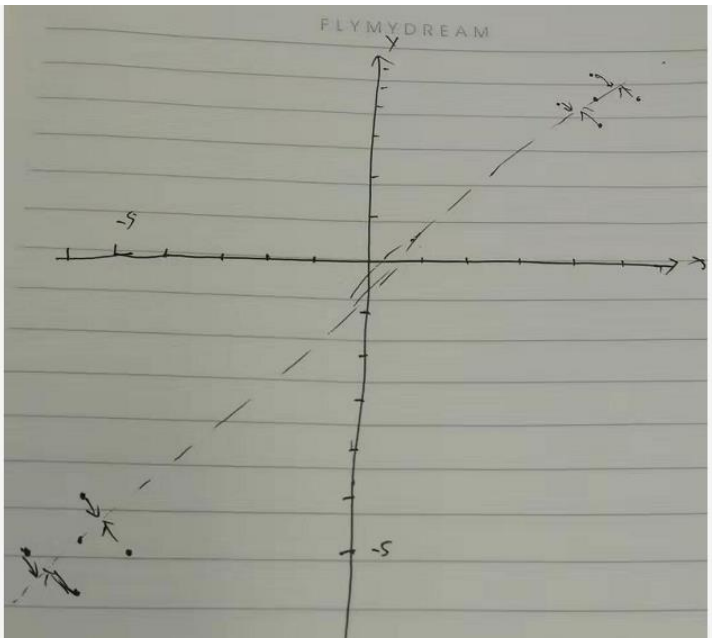
c) 然后计算特征值和特征向量 $\lambda=4$ ， $w=(1,1)^T$

d) 然后计算投影

$$w^TX=(1,1)\begin{Bmatrix}-5 & -5 & -4 & -5 & -6 & 5 & 5 & 6 & 5 & 4\\ -5 & -4 & -5 & -6 & -5 & 5 & 6 & 5 & 4 & 5\end{Bmatrix}$$

$$=\{\ -10\ ,-9\ ,-9\ ,-11\ ,-11,10,11,11,9\ ,9\ \}$$

e) 草图图示



---

5. （12 分）过拟合与欠拟合。

1) 什么是过拟合？什么是欠拟合？

2) 如何判断一个模型处在过拟合状态还是欠拟合状态？

3) 请给出 3 种减轻模型过拟合的方法。

i. 过拟合是指模型过于复杂但不具备泛化能力，在训练集合上表现好却在测试集合上

表现差。 欠拟合是指的模型简单，不能很好的拟合数据特征，在训练集合和测试集合上表现都不好。

ii. 如何判定模型 绘制模型复杂度-错误率的关系图，观察随着模型复杂度继续提升，如果训练误差减少而测试误差增大，说明模型过拟合，应当适当降低模型复杂度；如果训练误差和测试误差都在减少，说明模型欠拟合，应该继续增大模型复杂度。

iii. 减轻模型过拟合的方法：正则化技术；增加训练数据；添加随机因素；数据预处理和降维；提前终止迭代；决策树剪枝/集成学习……

---

6. （12 分）用逻辑回归模型（logistic regression model）解决K类分类问题，假设每个输入样本$x \in \mathbb{R}^d$的后验概率可以表示为：

$$P(Y=k|X=x)=\frac{\exp(w_k^T x)}{1+\sum_{l=1}^{K-1}\exp(w_l^T x)} \quad , \ k=1,\dots,K-1 \qquad (1)$$

$$P(Y=K|X=x)=\frac{1}{1+\sum_{l=1}^{K-1}\exp(w_l^T x)} \qquad (2)$$

其中w_k^T表示向量w_k的转置。通过引入$w_K=\vec{0}$，上式也可以合并为一个表达式。

1) 该模型的参数是什么？数量有多少？

2) 给定 n 个训练样本$\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$，请写出对数似然函数（log likelihood function）L的表达形式，并尽量化简。

$$L(w_1,\dots,w_{K-1})=\sum_{i=1}^{n}\ln P(Y=y_i|X=x_i)$$

3) 如果加入正则化项（regularization term），定义新的目标函数为：

$$J(w_1,\dots,w_{K-1})=L(w_1,\dots,w_{K-1})-\frac{\lambda}{2}\sum_{l=1}^{K}\|w_l\|_2^2$$

请计算 J 相对于每个w_k的梯度。

i. 模型的参数是$w_i^T \ , i=1,2\dots K-1$ 。参数数目共$(K-1)*d$ 个。

ii. 将y_i由标量扩展成 K 维度向量${y_i}^k$，其中 $if \ \ y_i=s \ , {y_i}^s=1 \ ; else \ , {y_i}^s=0$

$$L(w_1\dots,w_{K-1})=\sum_{i=1}^{n}\ln P(Y=y_i|X=x_i)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}\ln(P(Y=k|x=x_i)^{y_i^k})$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}(y_i^k\ (\ (w_k{}^T x)-y_i^k\ \ln(1+\sum_{i=0}^{k}\exp(w_k{}^T x))\ )$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}\ y_i^k(w_k{}^T x)-\ y_i^k\ \sum_{i=1}^{n}\sum_{j=1}^{k}\ln\left(1+\sum_{i=0}^{k}\exp(w_k{}^T x)\ \right)$$

$$=\sum_{i=1}^{n}(\sum_{j=1}^{k}y_i^k\ (w_k{}^T x)-ln\left(1+\sum_{i=0}^{k}\exp(w_k{}^T x)\ \right))$$

iii. 计算梯度：

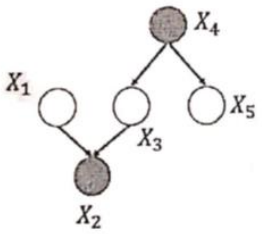
$$J(w) = L(w_1 \dots, w_{K-1}) - \frac{\lambda}{2}\sum_{l=0}^{k} ||w_k||_2^2$$

$$\frac{\partial J(w)}{\partial w_k} = \frac{\partial L(w_1 \dots, w_{K-1})}{\partial w_k} - \lambda w_k$$

$$= \sum_{i=0}^{n} x_i \left(y_i^k - \frac{\exp({w_k}^T x)}{1 + \sum_{i=0}^{k} \exp({w_k}^T x)} \right) - \lambda w_k$$

$$= \sum_{i=0}^{n} x_i \Big(y_i^k - P(Y = k|X = x_i) \Big) - \lambda w_k$$

7. （10 分）给定如下概率图模型，其中变量 X_2、X_4 为已观测变量，请问变量 X_1 和 X_5 是否独立？并用概率推导证明之。



i. 首先，在$x2, x4$已知的情况下，$x1$，$x5$是独立的。

ii. 证明：

$$p(x1, x2, x3, x4, x5) = p(x4) * p(x3|x4) * p(x5|x4) * p(x1) * p(x2|x3, x2)$$

$$p(x1, x2, x4, x5) = \sum_{x3} p(x1, x2, x3, x4, x5)$$

$$= \sum_{x3} p(x4) * p(x3|x4) * p(x5|x4) * p(x1) * p(x2|x3, x1)$$

$$= p(x4) * p(x5|x4) * p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)$$

$$p(x2, x4, x5) = \sum_{x1} p(x1, x2, x4, x5)$$

$$= p(x4) * p(x5|x4) \sum_{x1} p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)$$

$$p(x1|x2, x4, x5) = \frac{p(x1, x2, x4, x5)}{p(x2, x4, x5)} = \frac{p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)}{\sum_{x1} p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)}$$

$$p(x2, x4) = \sum_{x5} p(x2, x4, x5)$$

$$= p(x4) * \sum_{x5} p(x5|x2) * \sum_{x1} p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)$$

$$= p(x4) * \sum_{x1} p(x1) \sum_{x3} p(x3|x4)\, p(x2|x3, x1)$$

$$p(x1,x2,x4)=\sum_{x5}p(x1,x2,x4,x5)$$

$$=\ p(x4)*\sum_{x5}p(x5|x4)\,p(x1)\sum_{x3}p(x3|x4)\,p(x2|x3,x1)$$

$$=\ p(x4)*p(x1)\sum_{x3}p(x3|x4)\,p(x2|x3,x1)$$

$$p(x1|x2,x4)=\ \frac{p(x1,x2,x4)}{p(x2,x4)}=\frac{p(x1)\sum_{x3}p(x3|x4)\,p(x2|x3,x1)}{\sum_{x1}p(x1)\sum_{x3}p(x3|x4)\,p(x2|x3,x1)}$$

$$=p(x1|x2,x4,x5)$$

得证：$x1\perp x5\ |\ x2,x4$

---

8. （12 分）假设有 2 枚硬币，分别记为 A 和 B，以π的概率选择 A，以 1-π的概率选择 B，这些硬币正面出现的概率分别是p和q。掷选出的硬币，记正面出现为 1，反面出现为 0，独立地重复进行 4 次试验，观测结果如下：1，1，0，1。给定模型参数$\pi=0.4,p=0.6,q=0.5$，请计算生成该序列的概率，并给出该观测结果的最优状态序列。

这个问题其实不是隐马尔可夫模型，而是更简单的，事件流内部的事件之间是相互独立的。也就是不需要考虑转移概率$p(y_{t+1}|y_t)$，只有一个状态概率$p(y=A)=0.4\quad p(y=B)=0.6$ 和各自对应的发射概率，所以，计算$p(x)=p(x_1)*p(x_2)*p(x_3)*p(x_4)$：

$$p(1,1,0,1)=(\pi*p+(1-\pi)*q)*(\pi*p+(1-\pi)*q)$$

$$*\big(\pi*(1-p)+(1-\pi)*(1-q)\big)*(\pi*p+(1-\pi)*q)$$

$$=(0.4*0.6+0.6*0.5)^3*(0.4*0.4+0.6*0.5)$$

$$=0.54^3*0.46$$

$$=0.0765$$

因为事件流内部的事件相互独立，所以最优状态序列也是可以独立计算的：

$$p(y=A|x=1)=\frac{0.24}{0.54}$$

$$p(y=B|x=1)=\frac{0.3}{0.54}$$

$$p(y=A|x=0)=\frac{0.16}{0.46}$$

$$p(y=B|x=0)=\frac{0.3}{0.46}$$

因此最佳状态序列是 $\{\ B\ ,B\ ,B\ ,B\ \}$

---

9. （12 分）基于 AdaBoost 的目标检测需要稠密的扫描窗口并判断每个窗口是否为目标，请描述基于深度学习的目标检测方法，如 SSD 或 YOLO，如何做到不需要稠密扫描窗口而能发现并定位目标位置？

i. YOLO：

1) 将图像网格化，比如 7*7

2) 综合整个图片的信息，预测每个网格中物体边框的概率

ii. SSD

1) 网格式检测
2) Anchor 不同长宽比的物体框
3) 在不同尺度的特征图上检测

中国科学院大学

考试日期：2020.01.11

试 题 专 用 纸

课程编号：721004H

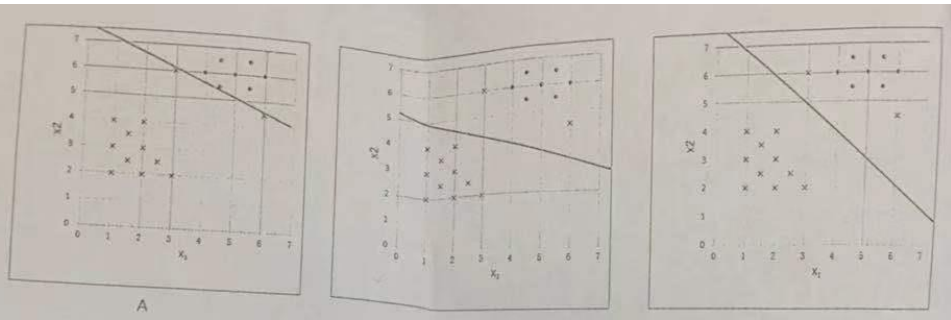
课程名称：模式识别与机器学习

任课教师：黄庆明等

姓名_____ 学号_____ 成绩_____

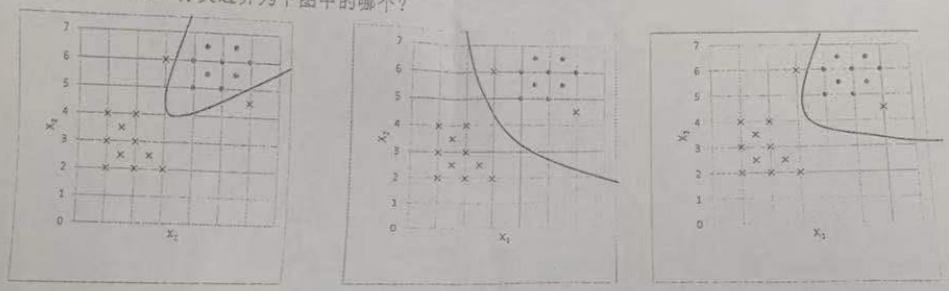
一、（16 分）选择题。（每个选项 2 分，请将答案写在答题纸上）

1. 基于二次准则函数的 H-K 算法较之于感知器算法的优点是哪个？
   A. 计算量小
   B. 可以判别问题是否线性可分
   C. 其解完全适用于非线性可分的情况
2. 在逻辑回归中，如果正则项取 L1 正则，会产生什么效果？
   A. 可以做特征选择，一定程度上防止过拟合
   B. 能加快计算速度
   C. 在训练数据上获得更准确的结果
3. 如果模型的偏差较高，我们如何降低偏差？
   A. 在特征空间中减少特征
   B. 在特征空间中增加特征
   C. 增加数据点
4. 假设采用正态分布模式的贝叶斯分类器完成一个两类分类任务，则下列说法正确的是哪个。
   A. 假设两类的协方差矩阵均为对角矩阵，则判别界面为超平面。
   B. 假设两类的协方差矩阵相等，则判别界面为超平面。
   C. 不管两类的协方差矩阵为何种形式，判别界面均为超平面。
5. 下列方法中，哪种方法不能用于选择 PCA 降维（K-L 变换）中主成分的数目K？
   A. 训练集上残差平方和随K发生剧烈变化的地方（肘部法）
   B. 通过监督学习中验证集上的性能选择K
   C. 训练集上残差平方和最小的 K
6. 考虑某个具体问题时，你可能只有少量数据来解决这个问题。不过幸运的是你有一个针对类似问题已经预先训练好的神经网络，请问可以用下面哪种方法来利用这个预先训练好的网络？
   A. 把除了最后一层外所有的层都冻住，重新训练最后一层
   B. 对新数据重新训练整个模型
   C. 只对最后几层进行调参(fine tune)
7. 如下图所示，假设该数据集中包含一些线性可分的数据点。训练 Soft margin SVM 分类器，其松弛项的系数为C。请问当$C \to 0$时，分类边界为下图中的哪个？



A　　B　　C

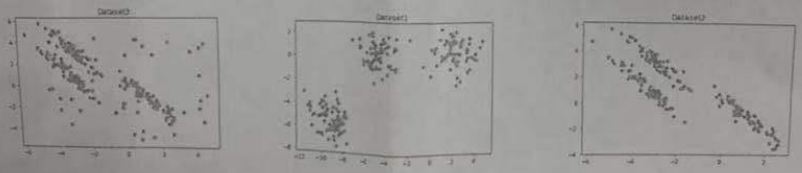
8. 如下图所示，假设该数据集中包含线性不可分的数据点。采用二次核函数训练 Soft margin SVM 分类器，请问当$C \to \infty$时，分类边界为下图中的哪个？



A　　B　　C

二、（6 分）请列举半监督学习对数据样本的三种基本假设。

三、（8 分）针对下图所示的三种数据分布，从 K 均值、GMM 和 DBSCAN 中分别选择最合适的聚类算法，并简述理由。



(a)　　(b)　　(c)

四、（12 分）对于具有类别标签的数据，采用 K-L 变换和 Fisher 线性判别分析两种方法对数据降维。

（1）简述这两种数据降维方法的基本过程。（8 分）

（2）这两种方法中哪种方法对分类更有效？并简述原因。（4 分）

五、（10 分）逻辑回归

（1）简述逻辑回归算法的原理。（4 分）

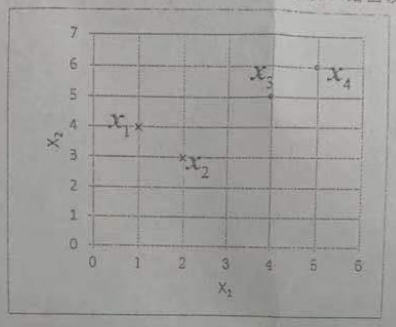
（2）如果使用逻辑回归算法做二分类问题得到如下结果，分别应该采取什么措施以取得更好的结果？并说明理由。（6 分）

（a）训练集的分类准确率 85%，验证集的分类准确率 80%，测试集的分类准确率 75%；
（b）训练集的分类准确率 99%，验证集的分类准确率 80%，测试集的分类准确率 78%；

六、（10 分）解释 AdaBoost 算法的基本思想和工作原理，并给出 AdaBoost 算法的伪代码。

七、（10 分）从特征提取的角度，分析深度卷积神经网络与传统特征提取方法（例如 Gabor 小波滤波器）的异同，并给出深度学习优于传统方法的原因。

八、（8 分）硬间隔支持向量机（Hard margin SVM）
如下图所示，一个数据集包含来自 2 个类别的 4 个数据点，在此集合上训练一个线性 Hard margin SVM 分类器。请写出 SVM 的形式化模型，并计算出该分类器的权重向量w和偏差b，给出该分类器的支持向量。



九、（10 分）拟利用贝叶斯判别方法检测 SNS 社区中不真实账号。设$Y=0$表示真实账号，$Y=1$表示不真实账号。每个用户有三个属性：X_1表示日志数量/注册天数，X_2表示好友数量/注册天数，X_3表示是否使用真实头像。已知$P(Y=0)=0.89$，$P(X_3=0|Y=0)=0.2$，$P(X_3=0|Y=1)=0.9$，且给定Y的情况下X_1、X_2的分布如下：

$P(X_1\|Y)$	$X_1 \le 0.05$	$0.05 < X_1 \le 0.2$	$X_1 \ge 0.2$
$Y=1$	0.8	0.1	0.1
$Y=0$	0.3	0.5	0.2
$P(X_2\|Y)$	$X_2 \le 0.1$	$0.1 < X_2 \le 0.8$	$X_2 \ge 0.8$
$Y=1$	0.7	0.2	0.1
$Y=0$	0.1	0.7	0.2

若一个账号使用非真实头像，日志数量与注册天数的比率为 0.1，好友数与注册天数的比率为 0.2，判断该账号是不是虚假账号。

十、（10 分）现装有红色球和白色球的两个盒子，盒子 1 中红球的比例为p，盒子 2 中红球的比例为q。我们以概率π选择盒子 1,概率$1-\pi$选择盒子 2，然后从盒子中有放回地取出一个小球，独立地重复进行 4 次试验，观测结果为：红，红，白，红。
*假定模型的参数初始值为$\pi^{(0)}=0.4$，$p^{(0)}=0.4$，$q^{(0)}=0.5$，*请写出 EM 算法迭代一次后p和q的值。（计算结果保留两位小数）