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）：

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

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）：

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

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： $w^* = S_w^{-1}(m_1 - m_2)$。

---

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。

ii. 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：

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

其中，$\hat{w}$和$\hat{w}_0$是分类超平面的参数，$\boldsymbol{\alpha}=[\alpha_1, \ldots \alpha_{|SV|}]$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$f(x;\boldsymbol{\alpha},\hat{w}_0)$在核特征空间的表达形式，进而证明：$f(x_{far};\boldsymbol{\alpha},\hat{w}_0)\approx \hat{w}_0$

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

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

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

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

证明:

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

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

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。

a) 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^TX 就是降维后特征

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

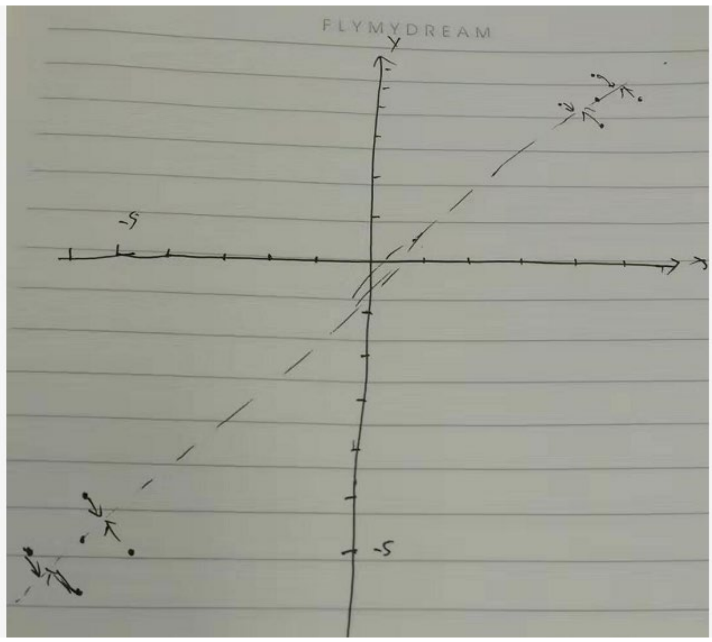
$$\rightarrow \begin{vmatrix} 4-\lambda & -4+\lambda \\ 250 & 254-\lambda \end{vmatrix} = 0$$

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

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

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

d) $W^TX = \{-\frac{10}{\sqrt{2}}, -\frac{9}{\sqrt{2}}, -\frac{9}{\sqrt{2}}, -\frac{11}{\sqrt{2}}, -\frac{11}{\sqrt{2}}, \frac{10}{\sqrt{2}}, \frac{11}{\sqrt{2}}, \frac{11}{\sqrt{2}}, \frac{9}{\sqrt{2}}, \frac{9}{\sqrt{2}}\}$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$w_i^T, i=1,2\dots K-1$ 。
   a) 如果认为x是增广的向量，参数数目共$(K-1)*d$ 个。
   b) 如果认为x是没有增广的向量，则是$(K-1)*(d+1)$个。

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

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

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

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

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

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

iii. 计算梯度:

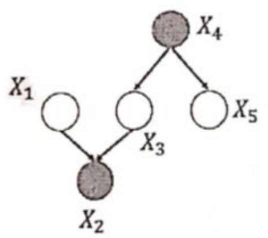
$$J(w) = L(w_1 \dots, w_{K-1}) - \frac{\lambda}{2}\sum_{l=0}^{k}||w_k||_2^2$$

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

得证：$x1 \perp x5 \mid x2, x4$

---

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$p(y = A) = 0.4 \ \ p(y = B) = 0.6$和各自对应的发射概率，所以，计算$p(x) = p(x_1) * p(x_2) * p(x_3) * p(x_4)$：

$$\begin{aligned} p(1,1,0,1) &= (\pi * p + (1-\pi) * q) * (\pi * p + (1-\pi) * q) \\ &\quad * \big(\pi * (1-p) + (1-\pi) * (1-q)\big) * (\pi * p + (1-\pi) * q) \\ &= (0.4 * 0.6 + 0.6 * 0.5)^3 * (0.4 * 0.4 + 0.6 * 0.5) \\ &= 0.54^3 * 0.46 \\ &= 0.0765 \end{aligned}$$

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