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。

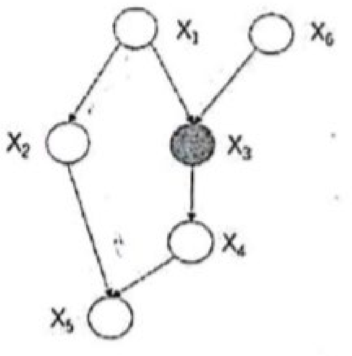
---

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 $w_{m+1,i}=\frac{w_{m,i}\exp\left(-\alpha_m y_i\emptyset_m(x_i)\right)}{Z_m}$

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。

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

$$C_1 = \frac{1}{N}\{(w_1-u_1)^T(w_1-u_1)\}；\ N = 4;$$

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

$$C_1 = \frac{1}{16}\begin{Bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{Bmatrix}$$

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

$$C_2 = \frac{1}{N}(w_2 - u_2)^T(w_2 - u_2)\,;N = 4$$

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

$$\Sigma_2 = \frac{1}{16}\begin{Bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{Bmatrix}$$

可见，$C_1 = C_2$ 又由于$p(w_1) = p(w_2)$，因此这是一个最小马氏距离分类器。

令$C = C_1 = C_2$；$C^{-1} = 4\begin{Bmatrix} 2 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{Bmatrix}$

$$\begin{aligned} d_1(x) - d_2(x) &= (u_1 - u_1)^T C^{-1} x - \frac{1}{2}u_1^T C^{-1} u_1 + \frac{1}{2}u_2^T C^{-1} u_2 \\ &= 8x_1 - 8x_2 - 8x_3 + 4 = 0 \end{aligned}$$

---

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 × 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。

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

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

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

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

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

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

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

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

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

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

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

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

---

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

$$令\theta=(\alpha,\mu,\Sigma)$$

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

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

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$\gamma(z_i^j) = 1\, if\;\; y_j = 1\, ; else\; \gamma(z_i^j) = 0$

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