一、（10 分）请简述最大似然估计和贝叶斯估计的区别

答案：

最大似然估计：把待估计的参数看作是确定的量，只是其取值未知。最佳估计就是使得产生已观测到的样本的概率最大的那个值。 （5 分）

贝叶斯估计：把待估计的参数看作是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正了对参数的初始估计值。 （5 分）

二、（10 分）从K个单高斯模型中采样，得到观测数据$\{x_1,\dots,x_N\}$。其中第k个单高斯模型服从分布$N(\mu_k,\Sigma_k)$，π_k表示观测数据属于第 k 个子模型的概率。请说出如何利用 EM 算法估计混合高斯模型参数，并说明得到的结果是否一定为最优解，若是，请简述理由，若不是，请简述可行的优化方法。

答案：

首先推导出似然函数；（2 分）

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \right\}$$

已知解一定满足如下等式：（3 分）

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \quad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^T$$

$$\pi_k = \frac{N_k}{N} \quad \gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\mu_j,\Sigma_j)}$$

则利用 EM 算法：（3 分）

a) 初始化 uk、Σk、Πk
b) E step:计算 r(znk)
c) M step:更新 uk、Σk、Πk

最终不一定是最优解。可以初始化几次不同的参数进行迭代，取结果最好的那次。（2 分）

三、（10分）给定两类别的数据集，其中正实例点为$X_1=(1,2)^T$，$X_2=(2,1)^T$；负实例为$X_3=(-1,1)^T$，$X_4=(-1,-1)^T$。请用感知机算法求出分类决策函数，并写出迭代步骤。感知机模型为$f(x)=sign(b+a^Tx)$，步长为1，初始解向量a_1=$(-1,0)^T$，$b_1=-1$（注：请按照实例下标的顺序对实例进行迭代）

答案：

首先对样本进行规范化和增广：（2分）

正实例：$y_1=(1,1,2)^T$，$y_2=(1,2,1)^T$

负实例：$y_3=(-1,1,-1)^T$，$y_4=(-1,1,1)^T$

下面进行迭代：初始时a_1=$(-1,-1,0)^T$

（1） ${a_1}^Ty_1=-2<0$，$a_2=a_1+y_1=(0,0,2)^T$ （1分）

（2） ${a_2}^Ty_2=2>0$，$a_3=a_2=(0,0,2)^T$ （1分）

（3） ${a_3}^Ty_3=-2<0$，$a_4=a_3+y_3=(-1,1,1)^T$ （1分）

（4） ${a_4}^Ty_4=3>0$，$a_5=a_4=(-1,1,1)^T$ （1分）

（5） ${a_5}^Ty_1>0$， $a_6=a_5=(-1,1,1)^T$ （1分）

（6） ${a_6}^Ty_2>0$， $a_7=a_6=(-1,1,1)^T$ （1分）

（7） ${a_7}^Ty_3>0$， $a_8=a_7=(-1,1,1)^T$ （1分）

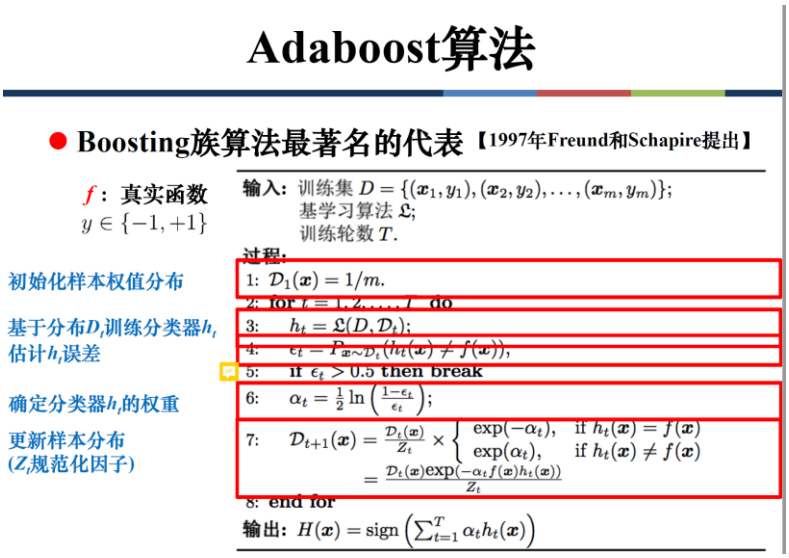
所以a=$(-1,1,1)^T$，分类决策函数为f(x)=sign(-1+x1+x2) （1分）

四、（10 分）（1）分别以 Adaboost 和 bagging 算法为例解释串/并行化方法原理（6 分）

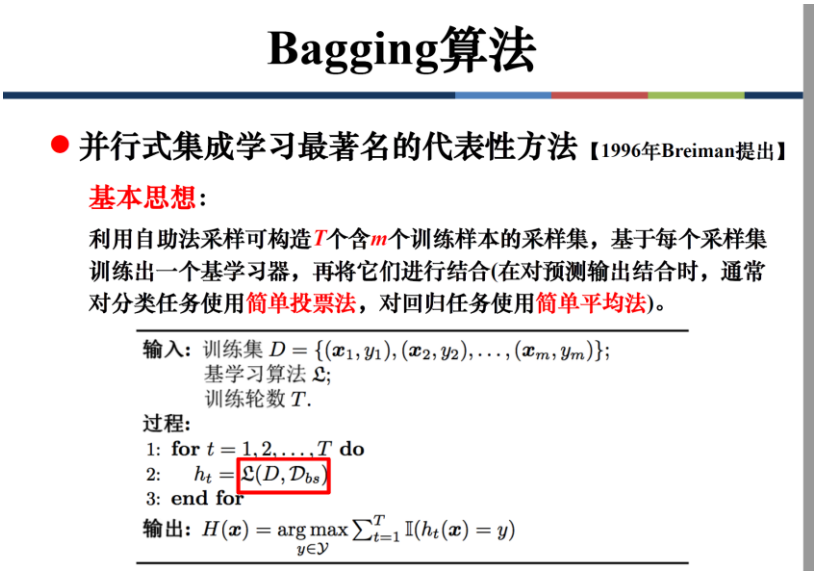
（2）请简述随机森林算法中如何引入样本扰动和属性扰动来实现基学习器的多样性（4 分）

答案：

（1） Adaboost：（3 分）



Bagging 算法：（3 分）



（2） 样本扰动：给定包含 N 个样本的数据集，经过 m 次有放回的随机抽样操作，得到 T 个含 m 个训练样本的采样集。（2 分）

属性扰动：对每个结点，从该结点的 d 个属性集合中随机选择包含 k 个属性的子集，再从这个子集中选择一个最优属性用于划分。（2 分）

五、（10 分）请简要说明支持向量机是如何使用核函数方法处理非线性数据的。

答案：

利用一个固定的非线性映射可以将数据映射到特征空间（2 分），在特征空间中学习的线性分类器等价于基于原始数据学习的非线性分类器（2 分）。在原问题中对数据的非线性映射等价于在对偶问题中将向量内积替换为核

函数（4 分），因此使用核函数的 SVM 模型可以实现非线性分类（2 分）。

六、（10分）PCA将D维数据集$\{x_n\}_{n=1}^N$降为M维($M < D$)，简述主成分分析（PCA）的最小均方误差思想**或**最大方差思想，并选择一种思想，推导 PCA，并给出 PCA 的简要计算步骤。

答案：

1. 简述思想（2 分）

   最小均方误差思想：使原数据与降维后数据在原空间中的重建误差最小。

   最大方差思想：使用较少的数据维度保留住较多的原数据特性。

2. 推导 PCA（4 分，合理即可）

   最小均方误差：

   1. 假设有一组 D 维的正交基$\{u_i\}_{i=1}^D$，每个数据可以表示为$x_n = \sum_i (x_n^T u_i)u_i$。令其中 M 个为 M 维子空间的基，每个数据可以表示为$x_n = \sum_i^M (x_n^T u_i)u_i + \sum_{M+1}^D (x_n^T u_i)u_i$，则降维后在原空间的表示为$\widetilde{x_n} = \sum_{M+1}^D (x_n^T u_i)u_i$。
   2. 重建误差定义为$J = \frac{1}{N}\sum_n |x_n - \widetilde{x_n}|$，变量为基向量
   3. $J = \sum_n \sum_{M+1}^D (x_n^T u_i - \widetilde{x_n} u_i)^2 = \sum_{M+1}^D u_i^T S u_i$，S为协方差矩阵。对其使用拉格朗日乘子法可以获得等式$Su_i = \lambda_i u_i$，也即u_i为特征向量。
   4. 将等式带入拉格朗日函数中，可以得知u_i对应于S的 D-M 个最小特征值的特征向量时，可以使得 J 最小。

   最大方差：

   1. 考虑一维情况，假设投影向量为u_1，投影后方差表示为$S' = u_1^T S u_1$，S为原样本方差
   2. 目标是$\max_{u_1} u_1^T S u_1, s.t. u_1^T u_1 = 1$
   3. 使用拉格朗日乘子法，得到等式$Su_i = \lambda_i u_i$
   4. 带入到拉格朗日函数中，可以得知u_1对应于协方差矩阵的最大特征值向量。多个向量的情况类似。

3. 简要步骤（4 分）

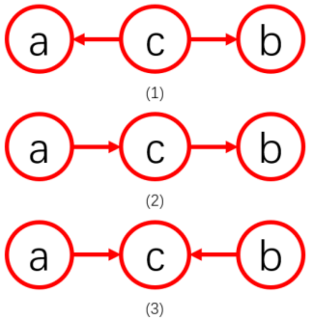
   ● 计算步骤

   ①计算给定样本$\{x_n\}, n = 1, 2, ..., N$的均值$\bar{x}$和协方差矩阵S;

   ②计算S的特征向量与特征值;

   ③将特征值从大到小排列，前M个特征值$\lambda_1, ..., \lambda_M$所对应的特征向量$u_1, ..., u_M$构成投影矩阵。

七、（10 分）请写出(1)，(2)，(3)三个概率图的联合概率分布，并判断每张图中的随机变量 a 与 b 是否独立。写出条件独立的定义，并判断每张图中在给定 c 的条件下，a 与 b 是否条件独立。



答案：

（1 分）$\mathrm{p(a,b,c)=p(a|c)p(b|c)p(c)}$

（1 分） $\mathrm{p(a,b)}=\sum_c p(a|c)\,p(b|c)p(c)$不一定等于 $\mathrm{p(a)p(b)}$，因此不独立

（1 分） $\mathrm{p(a,b|c)}=\frac{\mathrm{p(a,b,c)}}{\mathrm{p(c)}}=\mathrm{p(a|c)p(b|c)}$，因此条件独立

图(2)

（1 分）$\mathrm{p(a,b,c)=p}(a)\mathrm{p(c|a)p(b|c)}$

（1 分） $\mathrm{p(a,b)}=p(a)\sum_c p(c|a)p(b|c)$ 不一定等于 $\mathrm{p(a)p(b)}$，因此不独立

（1 分） $\mathrm{p(a,b|c)}=\frac{\mathrm{p(a,b,c)}}{\mathrm{p(c)}}=\frac{\mathrm{p}(a)\mathrm{p(c|a)p(b|c)}}{\mathrm{p(c)}}=\mathrm{p(a|c)p(b|c)}$，因此条件独立

图(3)

（1 分）$\mathrm{p(a,b,c)=p}(a)\mathrm{p(c|a,b)p(b)}$

（1 分） $\mathrm{p(a,b)}=p(a)p(b)$，因此独立

（1 分） $\mathrm{p(a,b|c)}=\frac{\mathrm{p(a,b,c)}}{\mathrm{p(c)}}=\frac{\mathrm{p}(a)\mathrm{p(c|a,b)p(b)}}{\mathrm{p(c)}}$不一定等于 $\mathrm{p(a|c)p(b|c)}$，因此不是条件独立

八、（20 分）（1）包含一层隐藏层的前馈神经网络如图 1 所示，其中给定训练集$D=\{(x_1,t_1),(x_2,t_2),\cdots,(x_N,t_N)\},x_i\in\mathbb{R}^D,t_i\in\mathbb{R}^K$。隐藏层激活函数为 h(x)，输出层激活函数为$\sigma$(x)。$y_n$表示第 n 个样本$x_n$对应的神经网络输出向量，$y_n=[y_{n1},\cdots,y_{nk},\cdots,y_{nK}]^T$，准则函数为$E_n(\mathrm{w})=\frac{1}{2}\sum_{\mathrm{k=1}}^{\mathrm{K}}\{\mathrm{y_{nk}-t_{nk}}\}^2$。网络包含 D 个输入神经元，K 个输出神经元，以及 M 个隐层神经元。试推导反向传播算法中对每一层权值参数（$\omega_{md}^{(1)}$ 与 $\omega_{km}^{(2)}$）的更新；（10 分）

（2）考虑只有一个神经元的多层神经网络，其中$x_{i+1} = \sigma(z_i) = \sigma(w_i x_i + b_i)\ (i \in [1,4])$，$C = Loss(x_5)$，$\sigma$表示 sigmoid 函数$f(x) = 1/(1+e^{-x})$，且$|w_i| < 1$。假设已知$\frac{\partial C}{\partial b_5}$，推导$\frac{\partial C}{\partial b_i}(i \in [1,4])$，并阐述当神经网络层数过深时，梯度消失的原因。 （10 分）

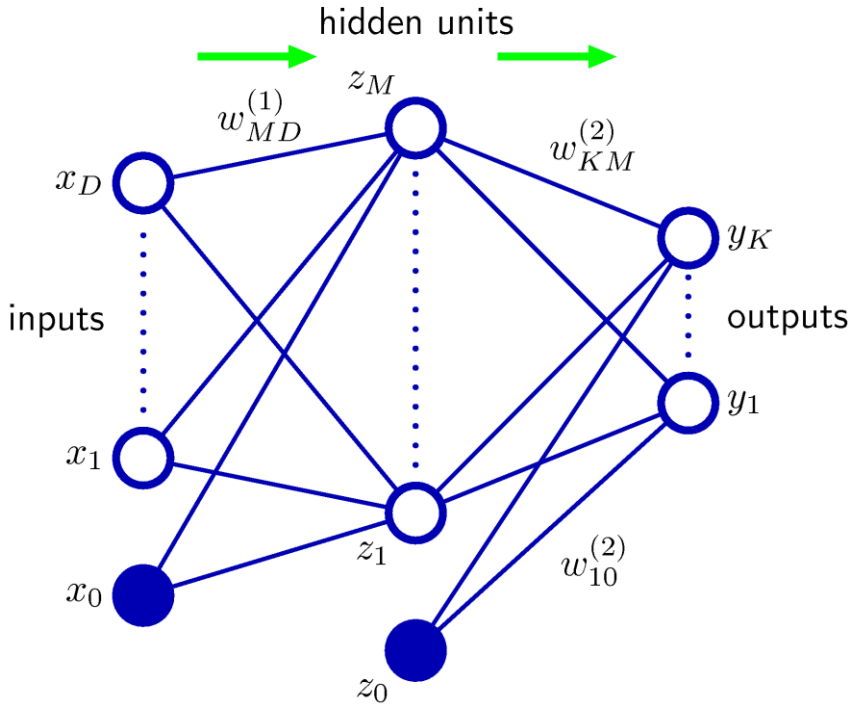


图 1

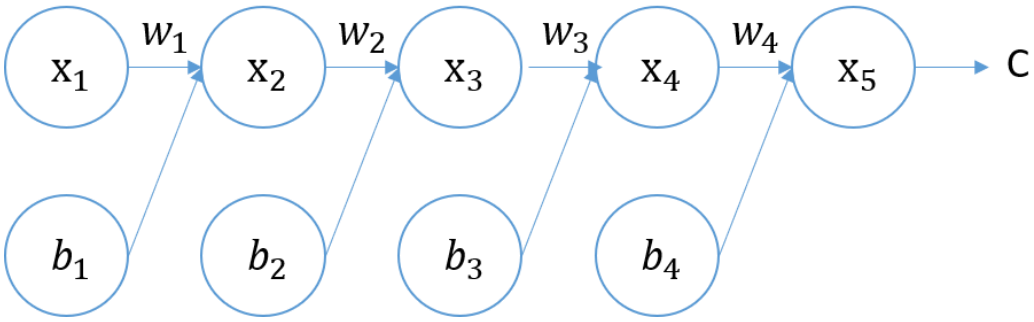


图 2

答案：

（1）网络前向传播如下： （2 分）

$$a_m = \sum_{d=0}^{D} \omega_{md}^{(1)} x_d$$

$$z_m = h(a_m)$$

$$a_k = \sum_{j=0}^{m} \omega_{km}^{(2)} z_m$$

$$y_{nk} = \sigma(a_k)$$

对于输出层：（2 分）

$$\delta_k = \frac{\partial E_n}{\partial a_k} = \frac{\partial E_n}{\partial y_{nk}} \cdot \frac{\partial y_{nk}}{\partial a_k} = (y_{nk} - t_{nk}) \cdot \sigma'(a_k)$$

对于隐藏层：（2 分）

$$\delta_m = \frac{\partial E_n}{\partial a_m} = \frac{\partial z_m}{\partial a_m} \cdot \sum_k \delta_k \cdot \frac{\partial a_k}{\partial z_m} = h'(a_m) \sum_k \delta_k \cdot \omega_{km}^{(2)}$$

因此，反向传播算法中对每一层权值参数更新值如下：（4 分）

$$\frac{\partial E_n}{\partial \omega_{km}^{(2)}} = \delta_k \cdot \frac{\partial a_k}{\partial \omega_{km}^{(2)}} = z_m \cdot (y_{nk} - t_{nk}) \cdot \sigma'(a_k)$$

$$\frac{\partial E_n}{\partial \omega_{md}^{(1)}} = \delta_m \cdot \frac{\partial a_m}{\partial \omega_{md}^{(1)}} = x_d \cdot h'(a_m) \sum_k \delta_k \cdot \omega_{km}^{(2)}$$

（2）可以推导出：

i=4: $\frac{\partial \mathrm{C}}{\partial \mathrm{b}_4} = \frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\frac{\partial \mathrm{x}_5}{\partial \mathrm{z}_4}\frac{\partial \mathrm{z}_4}{\partial \mathrm{b}_4} = \frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\sigma'(\mathrm{z}_4)$ （2 分）

i=3: $\frac{\partial \mathrm{C}}{\partial \mathrm{b}_3} = \frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\frac{\partial \mathrm{x}_5}{\partial \mathrm{z}_4}\frac{\partial \mathrm{z}_4}{\partial \mathrm{x}_4}\frac{\partial \mathrm{x}_4}{\partial \mathrm{z}_3}\frac{\partial \mathrm{z}_3}{\partial \mathrm{b}_3} =$

$\frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\sigma'(\mathrm{z}_4)\mathrm{w}_4\sigma'(\mathrm{z}_3)$ （2 分）

i=2: $\frac{\partial \mathrm{C}}{\partial \mathrm{b}_2} = \frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\frac{\partial \mathrm{x}_5}{\partial \mathrm{z}_4}\frac{\partial \mathrm{z}_4}{\partial \mathrm{x}_4}\frac{\partial \mathrm{x}_4}{\partial \mathrm{z}_3}\frac{\partial \mathrm{z}_3}{\partial \mathrm{x}_3}\frac{\partial \mathrm{x}_3}{\partial \mathrm{z}_2}\frac{\partial \mathrm{z}_2}{\partial \mathrm{b}_2} =$

$\frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\sigma'(\mathrm{z}_4)\mathrm{w}_4\sigma'(\mathrm{z}_3)\mathrm{w}_3\sigma'(\mathrm{z}_2)$ （2 分）

i=1: $\frac{\partial \mathrm{C}}{\partial \mathrm{b}_1} = \frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\frac{\partial \mathrm{x}_5}{\partial \mathrm{z}_4}\frac{\partial \mathrm{z}_4}{\partial \mathrm{x}_4}\frac{\partial \mathrm{x}_4}{\partial \mathrm{z}_3}\frac{\partial \mathrm{z}_3}{\partial \mathrm{x}_3}\frac{\partial \mathrm{x}_3}{\partial \mathrm{z}_2}\frac{\partial \mathrm{z}_2}{\partial \mathrm{x}_2}\frac{\partial \mathrm{x}_2}{\partial \mathrm{z}_1}\frac{\partial \mathrm{z}_1}{\partial \mathrm{b}_1} =$

$\frac{\partial \mathrm{C}}{\partial \mathrm{x}_5}\sigma'(\mathrm{z}_4)\mathrm{w}_4\sigma'(\mathrm{z}_3)\mathrm{w}_3\sigma'(\mathrm{z}_2)\mathrm{w}_2\sigma'(\mathrm{z}_1)$ （2 分）

由于$\sigma'(z) \leq \frac{1}{4}$且$|w| < 1$,则当层数过深时，$\frac{\partial C}{\partial b_1}$趋于 0，则梯度消失（2 分）

九、（10 分）请结合实际应用谈谈机器学习算法的局限性以及可能的解决方案

答案：合理即可