北京航空航天大学研究生课程试卷 **A**

2022－2023 学年第一学期期末试卷

学号_____ 姓名_____ 成绩_____

考试日期： 2022 年 12 月 22 日，上午 15:50－17:50

考试科目：《 机器学习 》（A 卷）

注意事项：1、请大家仔细审题

2、不能违反考场纪律

一.Minsky与Papert指出：感知机因为是线性模型，所以不能表示复杂的函数，如异或。异或函数输入$X_1=(1,-1)^T$，$X_2=(-1,1)^T$时，输出1；输入$X_3=(1,1)^T$，$X_4=(-1,-1)^T$时，输出-1。试着使用感知机算法求出分类决策函数，并写出迭代步骤，验证感知机为什么不能表示异或。感知机模型为$f(x)=\text{sign}(b+a^Tx)$，初始解向量$a_1=(0,0)^T$，$b_1=0$，梯度下降法的步长为1（注：请按照实例下标的顺序对实例进行迭代）。（10分）

答案：

首先对样本进行规范化和增广：

正实例：$Y_1=(1,1,-1)^T$，$Y_2=(1,-1,1)^T$

负实例：$Y_3=(-1,-1,-1)^T$，$Y_4=(-1,1,1)^T$（3 分）

下面进行迭代：初始$a_1=(0,0,0)^T$

$a_1^TY_1=(0,0,0)(1,1,-1)^T=0=0$，$a_2=a_1+Y_1=(1,1,-1)^T$（1.5 分）

$a_2^TY_2=(1,1,-1)(1,-1,1)^T=-1<0$，$a_3=a_2+Y_2=(2,0,0)^T$（1.5 分）

$a_3^TY_3=(2,0,0)(-1,-1,-1)^T=-2<0$，$a_4=a_3+Y_3=(1,-1,-1)^T$（1.5 分）

$a_4^TY_4=(1,-1,-1)(-1,1,1)^T=-3<0$，$a_5=a_4+Y_4=(0,0,0)^T$（1.5 分）

因为$a_5=a_1$，进入循环，所以感知机不能表示异或。（1分）

二.以下西瓜数据集包含 10 个训练样例，7 个测试样例，用来学习一棵能够预测没剖开的西瓜是不是好瓜的决策树。请根据下表，使用 ID3 算法建立决策树，并采用预剪枝策略进行剪枝。要求写出每个计算步骤，画出决策树的各层，指出叶子结点所代表的类别。注意：结点的类别标记为训练样本样例数最多的类别，如果样例数相等，则约定该结点类别标记为好瓜。（10 分）

表 1 训练样例

编号	色泽	根蒂	敲声	好瓜

1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	乌黑	蜷缩	浊响	是
6	青绿	稍蜷	浊响	是
7	乌黑	稍蜷	浊响	是
10	青绿	硬挺	清脆	否
14	浅白	稍蜷	沉闷	否
15	乌黑	稍蜷	浊响	否
16	浅白	蜷缩	浊响	否
17	青绿	蜷缩	沉闷	否

表 2 测试样例

编号	色泽	根蒂	敲声	好瓜
4	青绿	蜷缩	沉闷	是
5	浅白	蜷缩	浊响	是
8	乌黑	稍蜷	浊响	是
9	乌黑	稍蜷	沉闷	否
11	浅白	硬挺	清脆	否
12	浅白	蜷缩	浊响	否
13	青绿	稍蜷	浊响	否

答案：

计算根节点的信息熵为

$$Ent(D) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$（1 分）

若使用色泽划分， D^1，D^2，D^3分别表示色泽为青绿、乌黑、浅白。划分之后的信息熵为

$$Ent(D^1) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$Ent(D^2) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811$$

$$Ent(D^3) = -\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

可以计算出色泽的信息增益为

$$Gain(D，色泽) = 1 - \left(\frac{4}{10} * 1 + \frac{4}{10} * 0.811 + \frac{2}{10} * 0\right) = 0.276$$（1 分）

若使用根蒂划分，D^1，D^2，D^3分别表示色泽为蜷缩、稍蜷、硬挺。划分之后的信息熵为

$$Ent(D^1) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

$$Ent(D^2) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1$$

$$Ent(D^3) = -\left(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

可以计算出根蒂的信息增益为

$$Gain(D，根蒂) = 1 - \left(\frac{5}{10} * 0.971 + \frac{4}{10} * 1 + \frac{1}{10} * 0\right) = 0.115$$（1 分）

若使用敲声划分，D^1，D^2，D^3分别表示色泽为浊响、沉闷、清脆。划分之后的信息熵为

$$Ent(D^1) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$$

$$Ent(D^2) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$$

$$Ent(D^3) = -\left(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

可以计算出敲声的信息增益为

$$Gain(D，敲声) = 1 - \left(\frac{6}{10} * 0.918 + \frac{3}{10} * 0.918 + \frac{1}{10} * 0\right) = 0.174$$（1 分）

属性色泽的信息增益最大，选择它进行属性划分

色泽

|青绿 \乌黑 \浅白

{1, 6, 10, 17}好瓜 {2, 3, 7, 15}好瓜 {14, 16}坏瓜（0.5 分）

划分前{4, 5, 8}的样例被分类正确，验证集精度为 3/7=42.9%，使用色泽划分之后{4, 8, 11, 12}的样例被划分正确，验证集精度为 4/7=57.1%。因此使用色泽进行划分。（0.5 分）

然后决策树对D^1节点做进一步划分，该节点包含{1, 6, 10, 17}，属性集合为{根蒂, 敲声}。基于D^1计算出各属性的信息增益

$$Gain(D^1，根蒂) = 0.5$$

$$Gain(D^1，敲声) = 1$$（1 分）

属性敲声的信息增益最大，选择它进行属性划分

色泽

|青绿 \乌黑 \浅白

{1, 6, 10, 17}好瓜 {2, 3, 7, 15}好瓜 {14, 16}坏瓜

|

敲声

|浊响 \清脆 \沉闷

{1, 6}好瓜 {10}坏瓜 {17}坏瓜（0.5 分）

划分前{4, 8, 11, 12}的样例被分类正确，验证集精度为 4/7=57.1%，使用敲声划分之后{8, 11, 12}的样例被划分正确，验证集精度为 3/7=42.9%。因此不使用敲声进行划分。（0.5 分）

然后决策树将对D^2节点做进一步划分，该节点包含{2, 3, 7, 15}，属性集合为{根蒂, 敲声}。基于D^2计算出各属性的信息增益

$$Gain(D^2，根蒂) = 0.311$$

$$Gain(D^2，敲声) = 0.123$$（1 分）

属性根蒂的信息增益最大，选择它进行属性划分

色泽

|青绿 \乌黑 \浅白
{1,6,10,17}好瓜 {2,3,7,15}好瓜 {14,16}坏瓜
|
根蒂
|蜷缩 \稍蜷 \硬挺
{2，3}好瓜 {7，15}好瓜 {}好瓜（0.5分）

划分前{4,8,11,12}的样例被分类正确，验证集精度为4/7=57.1%，使用根蒂划分之后{4，8,11,12}的样例被划分正确，验证集精度为4/7=57.1%。因此不使用根蒂进行划分。（0.5分）

最终生成的决策树为

色泽
|青绿 \乌黑 \浅白
好瓜 好瓜 坏瓜（1分）

三. 说明K均值算法和混合高斯模型的异同点。简述如何使用EM算法估计混合高斯模型的参数。（10分）

答案：

相同点：K均值算法可看作高斯混合聚类在混合成分方差相等、且每个样本仅指派给一个混合成分时的特例。（2分）不同点：K均值用原型向量来刻画聚类结构不同，高斯混合聚类采用概率模型来表达聚类原型。（2分）

EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood. （1.5分）
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (9.23)$$ （1.5分）

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\,(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})\,(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \tag{9.27}$$

（3 分）

四.（1）简述支持向量机的基本思想，并写出模型表达式。说明支持向量机如何处理非线性可分的数据。（5 分）

（2）简述直推式支持向量机的基本思想，并写出模型表达式。说明直推式支持向量机如何避免穷举未标记样本的各种标记指派。（5分）

答案：

（1）SVM 寻找一个满足分类要求的超平面，并且使训练集中的点距离分类面尽可能的远，也就是寻找一个分类面使它两侧的空白区域（Margin）最大。（1.5 分）

$$\min_{w,b}\tfrac{1}{2}\|w\|^2$$
$$s.t.\ t_n(w^T x_n+b)\geq 1,\ n=1,...,N$$

（1.5 分）

利用一个固定的非线性映射可以将数据映射到特征空间，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。将线性支持向量机扩展到非线性支持向量机，只需要将线性支持向量机对偶形式中内积换成核函数。（2 分）

（2）T-SVM 同时利用标记和未标记样本，通过尝试将每个未标记样本分别作为正例和反例来寻找最优分类边界，来得到原始数据中两类样本的最大分类间隔。（1.5 分）

● **直推式支持向量机T-SVM**

$$\min_{\boldsymbol{w},b,\hat{\boldsymbol{y}},\boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l\sum_{i=1}^{l}\xi_i + C_u\sum_{i=l+1}^{m}\xi_i$$

未标记样本松弛变量

$$\text{s.t.} \ \ y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1-\xi_i, \ \ i=1,\ldots,l,$$

$$\hat{y}_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b) \geq 1-\xi_i, \ \ i=l+1,\ldots,m,$$

$$\xi_i \geq 0, \ \ i=1,\ldots,m,$$

未标记样本分类

其中，C_l和C_u分别表示已标记样本和未标记样本的惩罚因子，用于调整不同样本的权重，ξ为松弛变量，用于调整对错分样本的容忍程度

32 （1.5 分）

T-SVM 采用局部搜索迭代求近似解，通过局部搜索和调整指派为异类且可能错误的标记指派，使目标函数值不断下降。

局部搜索：x_i, x_j

异　　类：$y_iy_j<0$

可能错误：$\xi_i+\xi_j>2$

调　　整：互换标记 （2 分）

五. 在线性分类器设计的最小二乘准则中，b的选取具有多样性。试证明在二分类问题中，当$b=[\underbrace{N/N_1, N/N_1, N/N_1, \ldots}_{N_1\text{项}}, \underbrace{\ldots N/N_2, N/N_2, N/N_2}_{N_2\text{项}}]^T$（$N$为样本总数，$N_1$为第一类样本个数，$N_2$为第二类样本个数），此时基于最小二乘准则求解出的线性分类面等价于基于Fisher准则求解出的线性分类面。（10分）

答案：

证明如下：

首先对样本标签进行规范化，即$t_n=\begin{cases}\frac{N}{N_1}, 当x_n\in X_1\\ -\frac{N}{N_2}, 当x_n\in X_2\end{cases}$，最小二乘的优化目标$E$为：（1 分）

$$E=\frac{1}{2}\sum_{n=1}^{N}(w^Tx_n+w_0-t_n)^2 \tag{1}$$

E对w求导，得：（1 分）

$$\frac{\partial E}{\partial w}=\sum_{n=1}^{N}(w^Tx_n+w_0-t_n)x_n=0 \tag{2}$$

E 对 w_0 求导，得：（1 分）

$$\frac{\partial E}{\partial w_0}=\sum_{n=1}^{N}(w^T x_n+w_0-t_n)=0 \tag{3}$$

对（3）式展开进一步化解可得

$$\sum_{n=1}^{N}(w^T x_n)+Nw_0=0 \tag{4}$$

进一步求解（4）式可得：（1 分）

$$w_0=-\frac{1}{N}w^T\sum_{n=1}^{N}x_n \tag{5}$$

记 $m=\frac{1}{N}\sum_{n=1}^{N}x_n$，则 $w_0=-w^T m$，将 $w_0=-w^T m$ 代入到公式(2)中，得：（1 分）

$$\sum_{n=1}^{N}(w^T x_n+w^T m-t_n)x_n=0 \tag{6}$$

对（6）式进一步化解可得：

$$\sum_{n\in C_1}(x_n x_n^T-x_n m^T)w-\sum_{n\in C_1}t_n x_n+\sum_{n\in C_2}(x_n x_n^T-x_n m^T)w-\sum_{n\in C_2}t_n x_n=0 \tag{7}$$

对(7)式进一步化简可得：（2 分）

$$\{(\sum_{n\in C_1}x_n x_n^T+\sum_{n\in C_2}x_n x_n^T)-(N_1 m_1+N_2 m_2)m^T\}w=N(m_1-m_2) \tag{8}$$

在 Fisher 准则中，总类内离散度矩阵为：

$$S_w=\sum_{n\in C_1}(x_n-m_1)(x_n-m_1)^T+\sum_{n\in C_2}(x_n-m_2)(x_n-m_2)^T \tag{9}$$

对(9)式进一步化简可得：（1 分）

$$(\sum_{n\in C_1}x_n x_n^T+\sum_{n\in C_2}x_n x_n^T)=S_w+N_1 m_1 m_1^T+N_2 m_2 m_2^T \tag{10}$$

将(10)式代入到（8）式并进一步化简可得：

$$\{S_w+\frac{N_1 N_2}{N}(m_1 m_1^T+m_2 m_2^T-m_1 m_2^T-m_2 m_1^T)\}w=N(m_1-m_2) \tag{11}$$

又因为 $S_B=m_1 m_1^T+m_2 m_2^T-m_1 m_2^T-m_2 m_1^T$，因此（11）式可进一步化简为：（1 分）

$$S_w w+\frac{N_1 N_2}{N}S_B w=N(m_1-m_2) \tag{12}$$

又因为 $S_B w$ 与 m_1-m_2 同向，故（12）式等价于 $S_w w=\lambda(m_1-m_2)$，λ 为一个常数，因此基于最小二乘法最终求得的 $w\propto S_w^{-1}(m_1-m_2)$，故原问题得证。（1 分）

其他证明方法言之有理即可。

六. 存在一组二维样本：

$$X_1=[6,3]^T, X_2=[0,3]^T, X_3=[-3,-6]^T, X_4=[0,0]^T, X_5=[-3,0]^T$$

（1）使用主成分分析（PCA）的方法，将这五个点降维到一维空间。（6 分）

（2）简述使用 PCA 和 LDA（Fisher 准则）的基本思想和这两个方法的异同。（4 分）

答案：

（1）原始样本组成矩阵：

$$X=\begin{bmatrix}6 & 0 & -3 & 0 & -3\\ 3 & 3 & -6 & 0 & 0\end{bmatrix}$$

易求得以上 5 个样本的均值为 0，故协方差矩阵：（1 分）

$$C=\frac{1}{5}X\cdot X^T=\begin{bmatrix}\frac{54}{5} & \frac{36}{5}\\ \frac{36}{5} & \frac{54}{5}\end{bmatrix}$$

求解特征值为：（2 分）

$$\lambda_1=18, \lambda_2=\frac{18}{5}$$

使用第一个特征值 $\lambda_1=18$，求解特征向量为：（2 分）

$$v_1=\begin{bmatrix}-\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}\end{bmatrix}^T$$

样本点降至一维后，坐标分别为（1 分）

$$Y=v_1^T\cdot X=\begin{bmatrix}-\frac{9\sqrt{2}}{2} & -\frac{3\sqrt{2}}{2} & \frac{9\sqrt{2}}{2} & 0 & \frac{3\sqrt{2}}{2}\end{bmatrix}$$

（2）PCA：无监督降维方法，选取投影后方差最大的几个方向作为主成分。（2 分）

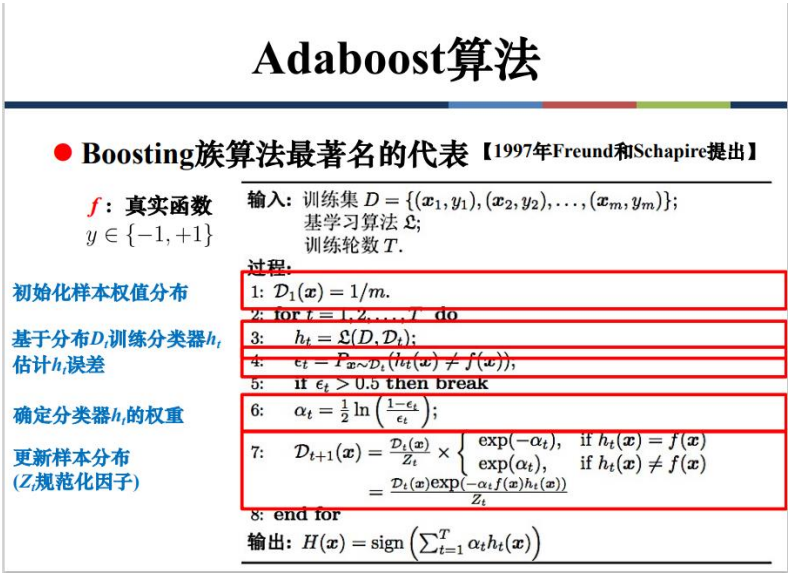
LDA：有监督降维方法，在投影过程中使类间方差大的同时类内方差小。（2 分）

七.（1）分别以 Adaboost 和 bagging 算法为例解释串/并行集成方法原理。（6 分）

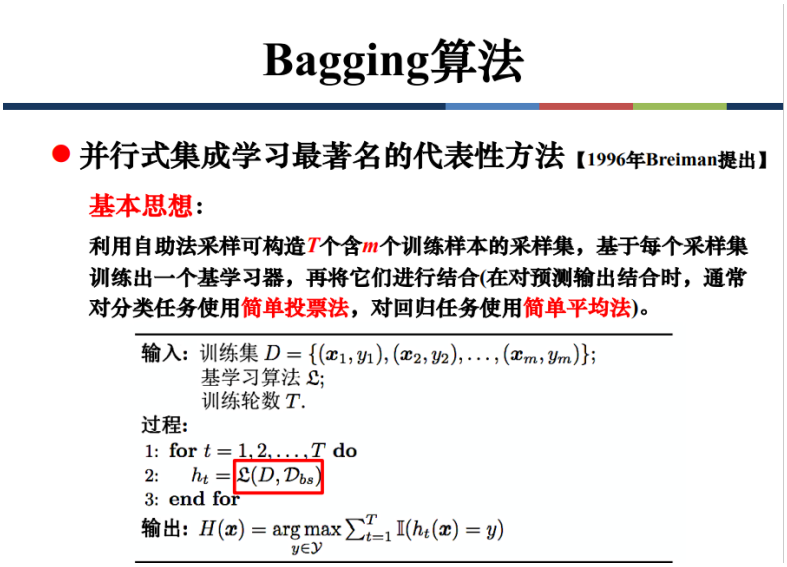
（2）请简述随机森林算法中如何引入样本扰动和属性扰动来实现基学习器的多样性。（4 分）

答案：

（1）Adaboost 算法：（3 分）



Bagging 算法：（3 分）



（2）样本扰动：给定 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：对每个节点，从该节点的 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）

八.（1）包含一层隐藏层的前馈神经网络如图 1 所示，给定训练集$D=\{(x_1,t_1),(x_2,t_2),\cdots,(x_N,t_N)\},x_i\in\mathbb{R}^D,t_i\in\mathbb{R}^K$。隐藏层包含$M$个神经元，激活函数为$\mathrm{h}(x)$，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$E_n(\mathrm{w})=\frac{1}{2}\sum_{k=1}^{K}\{y_{nk}-t_{nk}\}^2$。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）在训练深度神经网络时，经常采用正则化的方法来避免网络出现过拟合，请列举 4 种正则化的方法并简要说明其工作原理。（8 分）

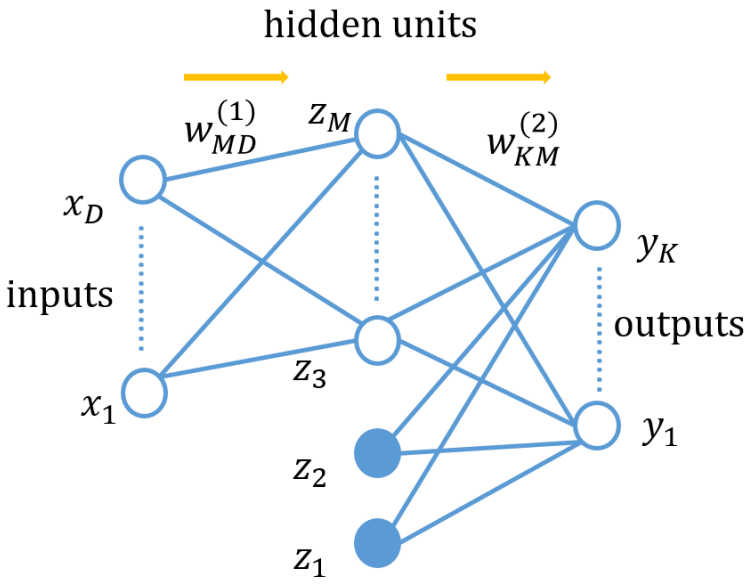


图 1

答案：

（1）网络的前向传播如下（2 分）：

$$a_m=\begin{cases}\sum_{d=1}^{D}w_{md}^{(1)}x_d & m\neq 1,2\\ a_m & m=1,2\end{cases}$$

$$z_m=\begin{cases}h(a_m) & m\neq 1,2\\ a_m & m=1,2\end{cases}$$

$$a_k=\sum_{m=1}^{M}w_{km}^{(2)}z_m$$

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

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）

$$\delta_m = \begin{cases} = \dfrac{\partial E_n}{\partial a_m} = \dfrac{\partial z_m}{\partial a_m} \cdot \sum_{k=1}^{K} \delta_k \dfrac{\partial a_k}{\partial z_m} = h'(a_m) \sum_{k=1}^{K} \delta_k \cdot w_{km}^{(2)} & m \neq 1,2 \\ = \dfrac{\partial E_n}{\partial a_m} = \dfrac{\partial z_m}{\partial a_m} \cdot \sum_{k=1}^{K} \delta_k \dfrac{\partial a_k}{\partial z_m} = \sum_{k=1}^{K} \delta_k \cdot w_{km}^{(2)} & m = 1,2 \end{cases}$$

因此，反向传播算法中对于每一层的权重更新参数如下：（4 分）

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

$$\frac{\partial E_n}{\partial w_{md}^{(1)}} = \delta_m \cdot \frac{\partial a_m}{\partial w_{md}^{(1)}} = x_d \cdot h'(a_m) \sum_{k=1}^{K} \delta_k \cdot w_{km}^{(2)}$$

（2）常用的 4 种正则化方法如下（答出 1 种正则化方法，给 2 分）：

1. L1 和 L2 正则化：是通过构建神经网络中优化参数的绝对值项之和项（L1）或者平方和项（L2），并将该项加入到损失函数中与任务损失联合进行优化，以保证网络中优化参数的稀疏性，进而避免过拟合。

2. Dropout：在神经网络前向传播的时候，让某个神经元的激活值以一定的概率 p 停止工作，这样可以使模型泛化性更强，因为结果不会太依赖某些局部的特征。

3. 数据增强：数据增强是指对输入样本数据（比如图像）进行翻转，裁剪，色彩变换等一些列操作而得到已有样本的相似样本作为训练数据训练网络，因为其丰富了数据分布，故而减少了网络的过拟合。

4. Early Stop: Early Stop 通过监控模型在一个额外的测试集上的表现来工作，当模型在测试集上的表现在连续的若干次（提前指定好的）迭代中都不再提升时它将终止训练过程。

九. 请结合机器学习算法设计一套火车站人脸识别系统，必要时可以画出流程图来进行辅助说明。（12分）

答案： 言之有理即可。