一、选择题（本题 30 分，每小题 3 分）

1. 关于有监督学习模型，下列说法中**错误**的是（ ）

A、回归和分类都是有监督学习问题。

B、回归模型一般采用平方误差损失函数。

C、CART可用于构建分类模型，但不可用于构建回归模型。

D、通过集成弱学习器可构建强学习器。

2. 关于SVM算法，下列说法中**错误**的是( )

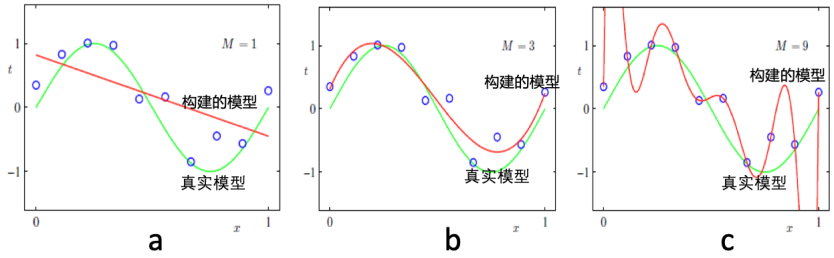
A、SVM的优化目标是最大化模型间隔

B、SVM属于结构风险最小化的模型

C、数据线性近似可分时，引入松弛变量有利于降低过拟合风险

D、SVM不属于核方法

3. 下图分别给出了采用多项式拟合的回归模型(图中M代表多项式次数)，关于这些模型描述下面说法**错误**的是（ ）



A、图（a）中多项式次数过低，模型出现欠拟合。

B、相比于图(a)和图（c），图（b）中模型具有更优的泛化性能。

C、图（c）中出现过拟合。

D、在多项式回归模型中采用正则化方法无法降低过拟合风险。

4 关于决策树构建，下列说法中**错误**的是( )

A、可通过剪枝降低模型过拟合风险

B、构建树时使得每个叶子节点仅含一类样本有助于提高预测性能

C、C4.5允许构建多叉树

D、信息增益、信息增益比和基尼指数均可用于评价和选择用于节点划分的特征

5 下列关于正则化的描述中，**错误**的是( )

A、正则化可以缓解模型过拟合问题　　B、线性回归模型中可采用L1或者L2正则化

C、正则化参数λ可采用交叉验证进行优化　　D、逻辑回归不能使用正则化

6 关于PCA，下列描述**正确**的是( )

A、对于数据矩阵**X**，当其特征数目大于样本数目时亦可采用PCA降维

B、PCA的目标函数是最小化投影残差，此目标等价于最大化主成分向量的方差

C、PCA求解可转化为特征值分解问题

D、PCA降维对奇异矩阵也适用

7 关于期望最大化(EM)算法，下列说法**错误**的是(　　)

A、EM算法是一种在潜变量(latent variables)存在条件下时的最大似然估计方法

B、EM通过对E-step和M-step交替迭代进行求解

C、在M-step中对模型参数进行估计

D、EM算法无法用于高斯混合模型求解

8 关于流形学习，下列说法**错误**的是（　）

A、Locally Linear Embedding通过最小化由近邻样本对每个样本的重构误差计算样本间的关联权重.

B、Laplacian Eigenmap方法的目标是最小化低维空间中样本间的距离

C、t-SNE优化目标是最小化原始高维空间中样本间相似度分布和低维空间中样本间相似度分布之间的距离。

D、ISOMAP采用欧式距离计算样本间的距离

9 下列说法中**正确**的是（　）

A、AdaBoost 抗过拟合能力很弱。

B、CART模型中基于具有3个及以上属性值的特征对样本进行划分时会产生多叉树

C、随机森林算法中的每棵树均是完全生长(fully grown)的树。

D、Bagging模型主要降低预测偏差。

10 关于损失函数，下列说法**错误**的是（　）：

A、逻辑回归中采用的是平方误差损失函数。

B、最小二乘回归中采用的是平方误差损失函数

C、SVM回归模型采用的是ε-不敏感损失函数。

D、在损失函数中引入正则化项有助于改善模型预测性能。

二、计算题（本题70 分）。

1. PCA（10 分）。给定由四个样本组成的矩阵X=[x_1， x_2， x_3， x_4]，其中 $x_1 = [0, 0]^T$，$x_2 = [1, 1]^T$，$x_3 = [2, 2]^T$，$x_4 = [-1, 1]^T$。计算降维到一维空间中的： (1)投影方向矢量u；(2)投影后的主成分向量z。(保留三位小数)

2. *k*-means（10 分）。给定训练集 $x_1 = [0, 1]^T$，$x_2 = [0, 2]^T$，$x_3 = [1, 2]^T$，$x_4 = [2, 1]^T$，$x_5 = [3, 0]^T$，$x_6 = [3, 1]^T$，若假定初始的类均值 $\mu_1 = [3, 1]^T$，$\mu_2 = [2, 1]^T$。采用 *k*-means 将其聚为两类，请写出两次迭代的计算过程和结果（采用欧式距离）。

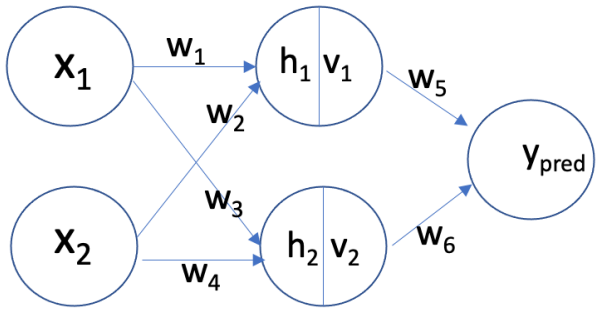
3. 决策树（10分）。 阿尔茨海默症(俗称老年痴呆症）与多种因素相关。下表给出了阿尔茨海默症患者和健康对照组的三个特征（β 淀粉样蛋白错误折叠、APOE4基因突变、性别）数据，试采用决

策树方法，构建阿尔茨海默症预测模型（以下每步计算请保留三位小数）

样本编号	β淀粉样蛋白错误折叠	APOE4基因突变	性别	阿尔茨海默症
1	是	有	男	是
2	是	有	女	是
3	否	无	男	否
4	否	无	女	否
5	是	有	男	否
6	是	有	男	是
7	是	无	女	是
8	否	无	男	否
9	否	有	女	否
10	否	无	男	是

(1) 求数据集的信息熵（3分）。
(2) 分别用β淀粉样蛋白错误折叠、APOE4基因突变和性别对数据集进行划分，计算相应的信息增益比，并给出最优特征（4分）。
(3) 采用基尼指数(Gini index)，对β淀粉样蛋白错误折叠和APOE4基因突变进行比较，判断出较优特征（3分）。
(提示：$\log_2(3)=1.5850$， $\log_2(5)=2.3219$)

4. 神经网络(15 分)。假设有如下结构的神经网络结构(如下图所示)： 其中隐层神经元采用sigmoid激活函数，其权重初始化为： $w_1=1$，$w_2=0.5$，$w_3=1$，$w_4=1$，$w_5=0.8$，$w_6=1$



(1) 给定样本$\mathbf{x}=[1,\ 0.5]^T$，试给出所有隐藏层节点的输入和输出（图中h_1和 h_2代表隐层神经元的输入，v_1和v_2代表隐层神经元的输出），以及预测值y_{pred}；(7分)
(2) 若$\mathbf{x}$对应的实际目标值y_{obs}为4，基于平方误差损失函数$L=0.5\times(y_{pred}-y_{obs})^2$，设置学习速率为0.1，Momentum参数$\rho=0.9$，试根据Momentum SGD法给出第一次更新后的w_1， w_2和w_5的值。（8分）

5. 混淆矩阵（10分）。新型冠状病毒肺炎（COVID-19）给人们的生活和全球经济带来了巨大的影响。在一项最新研究中，人们利用临床和医学影像数据，采用机器学习方法构建了一个能够预测疑似病人是否真正患有新冠肺炎的模型。在一个具有279个样本的测试集中，有112个新冠肺炎患者被预测为新冠肺炎患者，22个新冠肺炎患者没有被预测为新冠肺炎患者；27个不患有新冠肺炎的样本被预测为新冠肺炎患者，118个不患有新冠肺炎的样本被预测为不患有新冠肺炎。
根据上述数据：

(1) 给出混淆矩阵。(5分)

(2) 计算预测模型的敏感性(sensitivity)，特异性(specificity)，总体准确度(accuracy) (保留两位有效数字)。（5分）

6. SVM（15分）。某二维数据集如下图所示，其中圆圈为正样本(+1)，三角形为负样本(-1)。

(1) 假定采用线性核，试给出Hard SVM的支持向量与分类边界$w^T x + b = [w_1, w_2]\,[x_1, x_2]^T + b = 0$；（5分）

(2) 试给出$w = [w_1, w_2]^T$相对于分类边界$w^T x + b = 0$的几何意义以及与间隔(margin)的关系；（5分）

(3) 如果新增一个负样本$[x_1, x_2]^T = [5, 1]^T$，试重新给出Hard Linear SVM的支持向量与分类边界；（5分）