武汉大学 测绘学院
2021-2022学年度第一学期期末考试
《机器学习》课程试卷

题号	一	二	三	总分
分值	15	40	45	100

一、选择题（共 15 分，每题 1 分）

1. 设 ω_i 类样本服从一维正态分布，其类条件概率密度 $p(X|\omega_i)$ 服从正态分布，利用最大似然估计所估计的参数为（ ）。

A. μ、σ^2　　B. $p(\omega_i)$　　C. $p(X|\omega_i)$　　D. $p(\omega_i|X)$

2. 当假设空间 H 为二维空间中的矩形时，VC(H) 为（ ）。

A. 3　　B. 4　　C. 5　　D. 6

3. KNN（k 近邻）算法中的 k 指的是（ ）。

A. 样本特征维度　　B. 样本距离
C. 分类类别数　　D. 样本数量

4. 下列概念不正确的是（ ）。

A. 最大似然估计和贝叶斯估计都属于参数估计方法
B. 采用 0-1 损失函数时，最小风险贝叶斯决策等价于最小错误率贝叶斯决策
C. Bayes 决策的基本思想是：对判别问题，按类条件概率 $p(X|\omega_i)$ 最大作出决策
D. 朴素贝叶斯分类器采用“属性条件独立性假设”，即对已知类别，假设所有属性相互独立

5. 在现实应用中往往会遇到“不完整”的训练样本，这种未观测到的变量称为隐变量。对隐变量的估计通常采用哪种参数估计方法？（ ）

A. 朴素贝叶斯估计　　B. EM 算法
C. 最大似然估计　　D. 贝叶斯学习

6. 面对机器学习中的降维问题，在原始特征空间中从 d 维样本中找出提供最多信息的 d′ 维，并丢弃其他 d-d′ 维，这种做法称为（ ）。

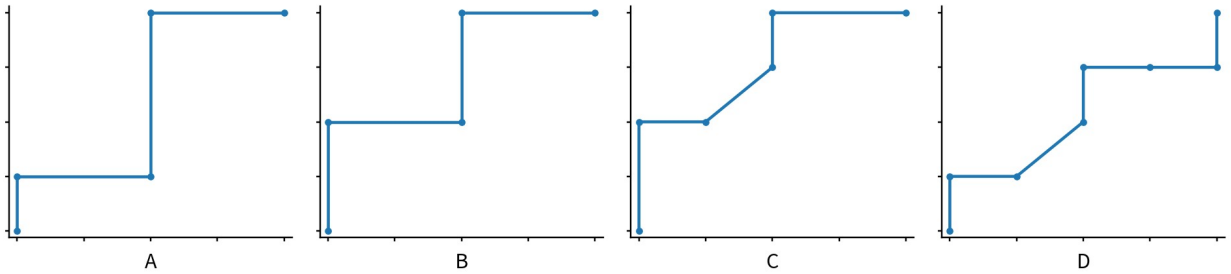
A. 样本提取　　B. 特征变换
C. 特征选择　　D. 特征采样

7. 用已经训练好的学习器对 8 个测试样本（4 个正例、4 个反例，即 $m^+=m^-=4$）进行预测，预测结果如下：

$(s_1, 0.80, +)$，$(s_2, 0.70, +)$，$(s_3, 0.56, -)$，$(s_4, 0.50, +)$
$(s_5, 0.50, -)$，$(s_6, 0.30, +)$，$(s_7, 0.20, -)$，$(s_8, 0.15, -)$

设横轴 x（假正例率）的单位刻度为 $1/m^-$，纵轴 y（真正例率）的单位刻度为 $1/m^+$。

那么 8 个测试样本的 ROC 曲线应选（　）。



8. 从偏差-方差分解来看，以下说法错误的是（　）。

A. Boosting 主要关注降低偏差，Bagging 主要关注降低方差
B. 随着训练程度加深，泛化误差由方差逐渐转变为偏差主导
C. 学习器欠拟合时，泛化误差由偏差主导
D. 神经网络过拟合时，泛化误差由方差主导

9. 已知正例点 $x_1=(1,2)^T$、$x_2=(2,3)^T$、$x_3=(3,3)^T$，负例点 $x_4=(2,1)^T$、$x_5=(3,2)^T$。最大间隔分类超平面的斜率为（　）。

A. 1　　B. 1/2　　C. 2　　D. -2

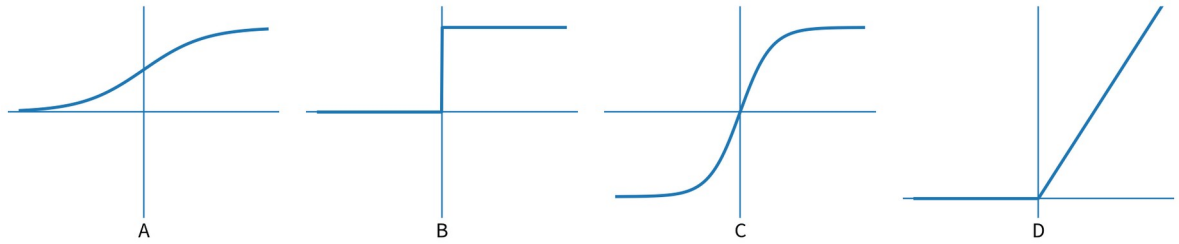
10. 影响梯度下降算法收敛效率的主要参数是（　）。

A. 学习率　　B. 权值　　C. 偏置值　　D. 测试误差

11. 一个卷积神经网络的输入为 32×32×3 的图像，第一个卷积层卷积核大小为 7×7，共 20 个卷积核，这一层总共含有（　）个参数（含偏置参数）。

A. 2960　　B. 2940　　C. 1000　　D. 980

12. 下列哪一个是 ReLU 激活函数（　）。



13. C4.5 决策树和 CART 决策树分别用（　）来进行属性划分。

A. 信息增益、信息增益率　　B. 信息增益率、基尼指数
C. 基尼指数、信息增益率　　D. 信息增益率、信息增益

14. 在决策树中可以通过（　）来避免过拟合。

A. 剪枝　　B. 缺失值处理　　C. 递归　　D. 信息熵

15. 在聚类算法中，可以用（　）度量输入样本的相似度。

A. DB 指数　　B. Rand 指数　　C. 参考模型　　D. 欧式距离

二、问答分析题（共 40 分，每题 8 分）

1. 假设三个判别函数如下，请依据下列三个判别函数回答问题：

$$g_1(x) = -x_1 - x_2 + 5$$

$$g_2(x) = -x_1 + 3$$

$$g_3(x) = -x_1 + x_2$$

（1）给出一对多情况下类别 1、2、3 的判别区域；判断并分析样本(1,2)及(4,5)的类别。（6 分）

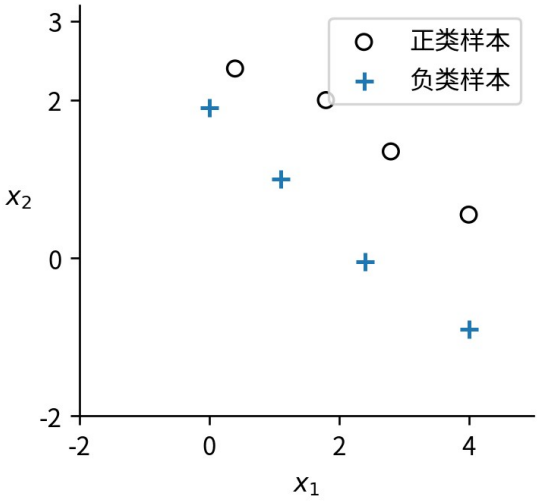
（2）若样本不能明确归类，请说明原因。（2 分）

2. 主成分分析（PCA）和线性判别分析（LDA）都是常用的降维方法，试回答：

（1）PCA 方法首先需要计算样本的协方差矩阵。对于 m 个 d 维样本 $(a_1,a_2,...,a_m)$，其协方差矩阵 Σ 的维度是多少？（1 分）

（2）如何使用 PCA 方法将样本维度降低至 d′ 维？简述核心步骤。（5 分）

（3）根据图 1 中两组样本的分布情况，分别用实线“——”和虚线“……”画出 PCA 和 LDA 的一般投影方向，并简述理由。（2 分）



3. 某数据集包含 1000 个样本，其中 500 个正例、500 个反例。请利用留出法和交叉验证法进行模型评估：

（1）简述如何利用留出法将其划分为包含 70%样本的训练集和 30%样本的测试集。（3 分）

（2）如何利用 5×5 交叉验证法进行模型评估？（3 分）

（3）简要说明两种方法的相同点和不同点。（2 分）

4. 下图为一个基本神经元及 BP 神经网络示意图，请回答：

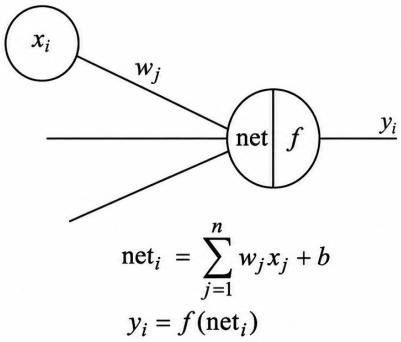


图 2　基本神经元

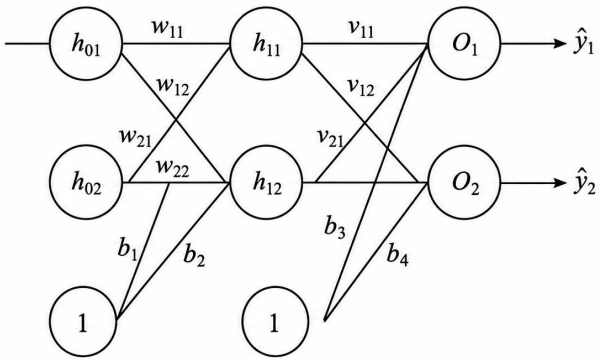


图 3 BP 神经网络（部分）

（1）如图 2、图 3 所示，每个神经元都经过 net 和 f 两个运算。其中 net 为输入信号的线性组合，f 称为激活函数。请总结激活函数的主要特点。（2 分）

（2）$\hat{y}_1$、$\hat{y}_2$ 分别为输出神经元 O_1 和 O_2 的输出值，其真实标注值为 y_1、y_2。请根据均方误差（MSE）给出该网络的损失函数 E。（2 分）

（3）简述 BP 神经网络的训练流程并指出主要参数。（4 分）

5. 用 DBSCAN 进行聚类时，有两个重要参数的设置对聚类效果有较大影响。图 4 展示了 320 个数据点的聚类结果，请回答：

（1）根据图 4 中的参数设置，说明参数 1 是什么、参数 2 是什么。（2 分）

（2）对比图 4 中的聚类结果，说明参数差异对 DBSCAN 聚类效果的影响，并分析原因。（2 分）

（3）给出 DBSCAN 算法聚类的基本步骤。（4 分）

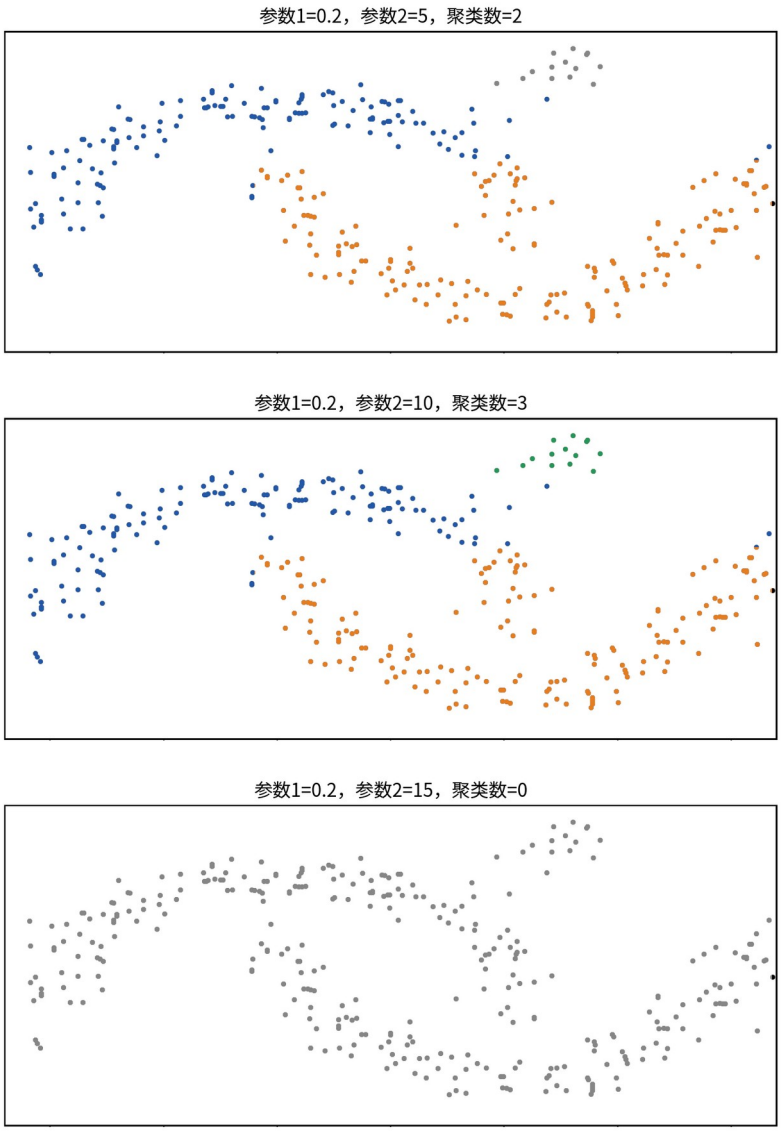


图 4 参数1=0.2，数据点320个时 DBSCAN 算法聚类结果

不同颜色表示不同聚类簇，灰色表示噪声点

三、分析计算题（共 45 分）

1. 假设在某个地区的组织识别中，正例 ω_1 和负例 ω_2 的先验概率分别为 $P(\omega_1)=0.8$、$P(\omega_2)=0.2$。现有一个待识别的组织，其观测值为 x。从类条件概率密度分布曲线上查得 $P(x|\omega_1)=0.25$、$P(x|\omega_2)=0.6$，并且已知 $\lambda_{11}=0$、$\lambda_{12}=6$、$\lambda_{21}=1$、$\lambda_{22}=0$。请对待识别组织 x 用下列两种方法进行分类：

（1）基于最小错误率的贝叶斯决策。（4 分）

（2）基于最小风险的贝叶斯决策。（6 分）

（3）请分析两种结果的异同及原因。（2 分）

2. 给定如下表所示的训练样本，三个弱学习器分别为：

$$h_1(x)=\begin{cases}1, & x<0.5\\ -1, & x>0.5\end{cases}\qquad h_2(x)=\begin{cases}-1, & x<4.5\\ 1, & x>4.5\end{cases}\qquad h_3(x)=\begin{cases}1, & x<7.5\\ -1, & x>7.5\end{cases}$$

请根据 AdaBoost 算法原理完成下表，并给出强分类器学习的计算过程（保留小数点后 4 位）：

（1）选出第一轮迭代中的最优弱分类器，并给出初始权重。（5 分）

（2）计算最优弱分类器的权重 α。（2 分）

（3）更新样本权重。（6 分）

（4）选出第二轮迭代中的最优弱分类器。（4 分）

样本点 X	0	1	2	3	4	5	6	7	8	9	错误数
实例 Y	1	1	1	-1	-1	1	1	1	-1	1	
第一轮分类结果											
h_1											
h_2											
h_3											
第二轮分类结果											
h_1											
h_2											
h_3											

3. 下表是某城市的一部分购房统计数据，现需据此构建购房决策树。

（1）请计算经验熵 H(D)，并说明信息增益和基尼指数分别用于何种决策树算法。（4 分）

（2）各属性的信息增益见表 1。请计算属性“面积”的经验条件熵和信息增益，确定相应决策树的根节点，并说明原因。（6 分）

（3）各属性的基尼指数见表 2。请分别计算属性“面积”和“单价=5 万”的基尼指数，确定相应决策树的最优切分点，并说明原因。（6 分）

注：上述计算采用 $\log_2$，并且需写出详细计算过程（保留小数点后 4 位）。

地段	近地铁	面积	单价（万）	是否购买
一环	是	60	5	是
一环	是	80	5	否
一环	否	60	4	是
一环	否	80	4	否
二环	是	60	5	是
二环	是	80	4	否
二环	否	60	3	是
二环	否	80	3	是
三环	是	60	3	是
三环	否	60	3	否

表 1 各属性信息增益表

	地段	近地铁	面积	单价
G(D,A)	0.21	0	?	0.096

表 2 各属性基尼指数表

属性	地段			近地铁	面积	单价		
	一环	二环	三环			5 万	4 万	3 万
基尼指数	0.467	0.4	0.4	0.48	?	?	0.419	0.45