湖南科技大学考试试题纸（A 卷）

（ 2021 - 2022 学年度 第 2 学期）

课程名称: 机器学习概论 **开课单位:** 数学学院 **命题教师:** 汤健

授课对象: 数学 学院 2019 年级 信息与计算科学 1-4 班

考试时量: 100 分钟 **考核方式:** 考查 **考试方式:** 开卷

审核人: _____ **审核时间:** _____年_____月_____日

一. 填空题（每空 2 分，共 30 分）

1. 根据训练数据是否拥有标记信息，学习任务可以大致划分为两大类：_____和_____；_____、_____是前者的代表，而_____是后者的代表；机器学习学的模型适用于新样本的能力，称之为_____。

2. _____和_____是科学推理的两大基本手段。

3. 当学习器把训练样本学得“太好了”的时候，很可能已经把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质，这样就会导致泛化性能下降，这种现象在机器学习中称之为_____，与之相对的是_____。

4. 线性判别分析（LDA）的思想非常朴素：给定训练样例集，设法将样例投影到一条直线上，使得_____的投影点尽可能接近，_____的投影点尽可能远离。

5. 多分类学习中最经典的拆分策略主要有哪三种：_____、_____和_____ 。

二. 计算题（第 1 题 16 分，第 2 题 24 分，共 40 分）

1. 给定数据集中的样本共分为 5 类，对样本 1 和样本 2，分别采用”一对一”（0 v 0）和”一对其余”（0 v R）两种策略进行分类，试写出分类结果。数据集为：

C1	C2	C3	C4	C5

对样本 1，采用 ”一对一”（0 v 0）策略，该样本属于_____ ；（8 分）

用于训练的两类样例		训练	分类器	输入样本 1	预测结果
+	−				
C1	C2	→	F1	→	+
C1	C3	→	F2	→	−
C1	C4	→	F3	→	−
C1	C5	→	F4	→	−
C2	C3	→	F5	→	+
C2	C4	→	F6	→	−

C2	C5	→	F7	→	−
C3	C4	→	F8	→	+
C3	C5	→	F9	→	−
C4	C5	→	F10	→	−

对样本 2，采用 “一对多”（O v R）策略，该样本属于_____；（8 分）

用于训练的两类样例		训练	分类器	输入样本 2	预测结果
+	−				
C1	C2, C3, C4, C5,	→	G1	→	+
C2	C1, C3, C4, C5	→	G2	→	−
C3	C1, C2, C4, C5	→	G3	→	−
C4	C1, C2, C3, C5	→	G4	→	−
C5	C1, C2, C3, C4	→	G5	→	−

2. 根据样本集 D(P86 表 4.4)上的属性“敲声”数据，

① . 写出该属性上无缺失值的样例子集$\tilde{D}$；（6 分）

② . 计算该样例子集$\tilde{D}$的信息熵（保留到小数点后三位）；（6 分）

③ . 令$\tilde{D}^1$、$\tilde{D}^2$与$\tilde{D}^3$分别表示在属性“敲声”上取值为“浊响”、“沉闷”以及“清脆”的样本子集，分别计算该三个样本子集的信息熵（保留到小数点后三位）；（9 分）

④ . 计算样本子集$\tilde{D}$上属性“敲声”的信息增益（保留到小数点后三位）。（3 分）。

三. 证明题（第小题 10 分，共 30 分）

1. 对于图 5.7，试推导出 BP 算法中的更新公式（5.12）和（5.13）.
2. 试证明样本空间中任意点 x 到超平面(w,b)的距离公式（6.2）.
3. 书本 P59,为什么说最大化（3.25 式）等价于最小化（3.27 式），试证明之。