第一章 概述

一、填空题

1 将模式识别任务看做 1 个输入输出系统$y = f(\boldsymbol{x})$，则对于分类任务，y通常为（类别标签）；对于回归任务，y 通常为（待预测变量）。

2 利用一个线性分类器对 iris 数据库进行分类，其样本空间是（所有可能的三类鸢尾花样本）的集合，维度是（4）；其假设空间是（线性判别函数）的集合，其参数空间维度是（5）；其类别标签集合是（'setosa','versicolour','virginica'）。

二、判断题

3 一个数据集的特征维度总是小于该数据集的样本数量。（×）

4 模型容量大小与训练集样本数量无关。（√）

5 明日降水概率预测属于回归问题。（√）

6 班级内自由分组讨论不属于分类问题。（√）

7 驯兽师用鞭打和喂食的方式教导老虎跳火圈属于有监督学习。（×）

三、选择题

8 处理二维数据的单个神经元模型的参数个数是：（A）

A. 3 个　　B. 2 个　　C. 1 个　　D. 不确定

9 从当前考场的所有考生中，找出作弊的学生，该任务属于：（C）

A. 分类　　B. 聚类　　C. 异常检测　　D. 以上均是

10 以下方法中不属于主流机器学习方法的是：（C）

A. 有监督学习　　B. 无监督学习　　C. 随机学习　　D. 强化学习

四、简答题

11 什么是分类任务？分类任务有哪些实际的应用？

分类任务是："根据观测数据判断一个对象类别"，实际应用包括人脸识别、医疗诊断、垃圾邮件过滤、无人驾驶中的交通信号识别……

12 什么是回归任务？回归任务有哪些实际的应用？

回归任务是："根据观测数据去预测对象的某个连续属性的数值"，实际应用包括天气预报、商品价格估计、目标坐标预测、年龄估计的……

13 分类任务和回归任务的区别是什么？

从输入输出系统的角度看，分类任务的输出是类别标签，通常为离散值；回归任务的输出是某种带预测变量，通常为连续值。

14 什么是聚类任务？请列举几个聚类的实际应用。

聚类任务是："将观测数据分为由相似对象组成的多个类别"，实际应用包括互联网客户分析、图像分割、网站资源管理、故障诊断……

15 什么是异常检测任务？异常检测的实际应用是什么？

异常检测任务是："找出不属于某个类别的异常对象"，实际应用包括信用卡异常检测、账户异常登录检测、智能视频监控中的异常目标与异常行为识别……

16 人脸图像伪造属于分类任务还是回归任务，为什么？

人脸伪造图像属于回归任务，因为其生成的图像可以看做是一个高维连续值变量。

17 Iris 数据集有哪几个类别？每个样本都包含那些特征？

Iris 数据库有 setosa，versicolor，virginica3 个类别，每个样本包含花萼长度、花萼宽度、花瓣长度、花瓣宽度 4 个特征。

18 模式识别方法可以分为哪三种思路？请分别举例说明！

基于知识的方法，例如基于动力学方程预测乒乓球的状态；

基于经验的方法，例如基于专家系统为商品质量分类；

基于学习的方法，例如基于人脸大数据实现人脸识别。

19 现代机器学习的三种主流方法是什么，各自的定义和特点是什么？

有监督学习，利用训练数据中给出的"标准答案"来学习，精度高，速度快，人工成本高；

无监督学习，没有标准答案，从数据分布中寻找答案，精度低，速度较快，人工成本低；

强化学习，没有标准答案，但有奖惩信号，从探索中寻找答案，精度较高，

速度较慢，人工成本低

20 如果一个模型在训练集上表现良好，但在测试集上表现不佳，试说明该模型的容量与问题复杂度之间的关系。

该模型的容量可能大于问题复杂度，导致了过拟合现象。

五、计算(画图)题

21 下面是一个房屋价格预测数据集的部分数据，该任务是分类任务还是回归任务？是有监督还是无监督？试写出每个样本的特征向量？特征空间的维度是多少？

编号	占地面积/m^2	距市中心距离/km	地下室面积/ m^2	房屋价格/万元
1	100	12	3	81.5
2	120	5	6	126
3	90	18	5	76.2
4	150	3	10	165.3
5	200	20	0	161
6	160	9	8	152.6

该任务是回归任务,属于有监督学习,1 号样本的样本特征向量为$[100,12,3]^T$,依此类推可以写出每个样本的特征向量。特征空间维度为 3.

22 请画出有监督机器学习系统的学习阶段的系统框图