中国科学院大学

试 题 专 用 纸

所属学期：2023 学年度秋季
课程编号：180206081100M1001H-01
180206081100M1001H-02
课程名称：模式识别
任课教师：向世明、孟高峰、张煦尧、张燕明

姓名_____ 学号_____ 成绩_____

1. （10 分）对一个 c 类分类问题，假设各类先验概率为 $P(\omega_i), i=1,\ldots,c$，条件概率密度为 $P(\mathbf{x}\mid\omega_i), i=1,\ldots,c$（这里 $\mathbf{x}$ 表示特征向量），将第 j 类模式判别为第 i 类的损失为 λ_{ij}。

（1）（5 分）请写出贝叶斯最小风险决策和最小错误率决策的决策规则；

（2）（5 分）引入拒识（表示为第 c+1 类），假设决策损失为

$$\lambda_{ij}=\begin{cases}0, & i=j\\ \lambda_r, & i=c+1\\ \lambda_s, \text{otherwise}\end{cases}$$

请写出最小损失决策的决策规则（包括分类规则和拒识规则）。

2. （10 分）表示模式的特征向量 $\mathbf{x}\in R^d$，对一个 c 类分类问题，假设各类先验概率相等，每一类条件概率密度为高斯分布。

（1）（3 分）请写出类条件概率密度函数的数学形式；

（2）（3 分）请写出在下面两种情况下的最小错误率决策判别函数：(a)类协方差矩阵不等；(b)所有类协方差矩阵相等。

（3）（4 分）在基于高斯概率密度的二次判别函数中，当协方差矩阵为奇异时，判别函数变得不可计算。请说出克服协方差矩阵奇异的方法。

3. （10 分）在 d 维特征空间中估计概率密度函数 $p(\mathbf{x})$ 有不同方法。

（1）（3 分）说明概率密度估计的参数法和非参数法各有什么特点；

（2）（3 分）说明 Parzen 窗估计的基本原理；

（3）（4 分）写出球形窗函数，以及采用球形窗函数情况下的 Parzen 窗估计概率密度函数；说明超球半径对密度估计的影响。

4. （共 10 分）广义线性判别函数。

（1）（4 分）设计一个线性判别函数 g(x)解决逻辑或(logic OR)问题，即 g(x)将样本(0, 1)，(1, 0)，(1, 1)分为第一类，将样本(0, 0)分为第二类；并画出 g(x)的决策面。

（2）（6 分）设计一个判别函数 g(x)解决异或问题（exclusive OR, XOR），即 g(x)将样本(0, 1)，(1, 0)分为第一类，将样本(0, 0)，(1, 1)分为第二类；进一步，将 g(x)表示为广义线性判别函数的形式。

5. （共 10 分）线性判别分析（Linear Discriminant Analysis, LDA）是一种经典的监督降维方法，在实际中广泛使用。

（1）（2 分）说明 LDA 的算法思想；

（2）（4 分）对于两类问题，给出样本的类内散度矩阵和类间散度矩阵的数学表示，给出 LDA 的优化目标；

（3）（4 分）对于两类问题，给出求解 LDA 优化问题的推导过程。

6. （15 分）神经网络。

(1) （5 分）试述将线性函数$f(x) = w^T x$用作神经元激活函数的缺陷；

(2) （5 分）试画出 sigmoid 函数和 ReLU 函数，并分析二者作为神经元激活函数的优劣；

(3) （5 分）设计一个用于图像分类问题的卷积神经网络，给出该网络的详细网络结构和参数设置，并描述各网络模块的功能。

7. （10 分）数据聚类。给定 5 个样本的集合$\{x_1, x_2, x_3, x_4, x_5\}$，样本之间的欧式距离由如下矩阵 D 表示：

$$D = [d_{ij}]_{5\times 5} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 7 & 2 & 9 & 3 \\ 7 & 0 & 5 & 4 & 6 \\ 2 & 5 & 0 & 8 & 1 \\ 9 & 4 & 8 & 0 & 5 \\ 3 & 6 & 1 & 5 & 0 \end{bmatrix} \end{array}$$

其中，d_{ij}表示第i个样本与第j个样本之间的欧式距离。

(1)（5 分）请给出簇D_i与簇D_j最小距离的定义公式，并计算簇$\{x_1\}$与簇$\{x_3, x_5\}$之间的最小距离；

(2)（5 分）采用最小距离，对这 5 个样本进行分级聚类，并画出最终聚类结果的系统树。

8. （10 分）决策树。

（1）（3 分）决策树方法 ID3 和 C4.5 的主要区别是什么？

（2）（3 分）如何防止决策树过拟合？

（2）（4 分）请阐述随机森林（Random Forests）的原理和优势。

9. （15 分）支撑向量机。假设给定一个特征空间上的训练数据集$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$，其中，$x_i \in R^n, y_i \in \{+1, -1\}, i = 1,2, \dots, N, x_i$是第$i$个特征向量，$y_i$为$x_i$的类别标记。假设上述数据为线性可分的，利用硬间隔线性支撑向量机对上述数据进行分类。

(1)（5 分）试写出上述硬间隔线性支撑向量机的原问题和对偶问题；

(2)（5 分）假设$\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)$是上述对偶问题的解。请写出 SVM 对应的分类决策函数；

(3)（5 分）请指出$\alpha_i^* (1 \le i \le N)$满足什么条件时，对应的特征向量$x_i$为支撑向量？并指出支撑向量的几何含义。