中国科学院大学

试 题 专 用 纸

所属学期：2017-2018 学年秋季第一学期
课程编号：251M1001H
课程名称：模式识别
任课教师：刘成林、向世明、张煦尧

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 $P(\mathbf{x}|\omega_i), i=1,\ldots,c$（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）在二维特征空间中，两个类别的概率密度为高斯分布（正态分布），参数分别为 $\mu_1=(1,0)^T$，$\mu_2=(-1,0)^T$，$\Sigma_1=\Sigma_2=\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}$，先验概率 $P(\omega_1)=P(\omega_2)=0.5$。

（1）（6 分）请给出分类误差最小的贝叶斯决策的决策面函数，并写出贝叶斯错误率（写成积分形式即可）；

（2）（4 分）当 $\lambda_{11}=\lambda_{22}=0, \lambda_{12}=2\lambda_{21}$，请给出损失最小的贝叶斯决策的决策面函数。

3. （15 分）特征空间中概率密度的非参数估计近似为 $p(\mathbf{x})=\dfrac{k/n}{V}$，其中 V 为 $\mathbf{x}$ 周边邻域的体积，k 为邻域内样本数，n 为总样本数。基于此定义，

（1）（5 分）请说明 Parzen 窗估计和 k-近邻（k-NN）估计的区别。

（2）（5 分）设一维特征空间中的窗函数 $\varphi(u)=\begin{cases}1, & |u|<1/2\\ 0, & \text{otherwise}\end{cases}$，现有 n 个样本 x_i，$i=1,\ldots,n$，采用宽度为 h_n 的窗函数，请写出概率密度函数 $p(x)$ 的 Parzen 窗估计 $p_n(x)$；

（3）（5 分）给定一维空间中的三个样本点{-1,0,2}，请写出概率密度函数 $p(x)$ 的最近邻（1-NN）估计并画出概率密度函数曲线图。

4. （共 11 分）

（1）（6 分）对多类分类可采用 one-vs-all 技巧构建 c 个线性判别函数：$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + b_i$，$i = 1, 2, \ldots, c$，并假定决策规则为：“对 $j \neq i$，如果 $g_i(\mathbf{x}) > g_j(\mathbf{x})$，$\mathbf{x}$ 则被分为 ω_i 类”。现有一个二维空间中的三类分类问题，其判别函数分别为 $g_1(\mathbf{x}) = -x_1 + x_2$；$g_2(\mathbf{x}) = x_1 + x_2 - 5$；$g_3(\mathbf{x}) = -x_2 + 1$。请画出分类决策面。

（2）（5 分） 请简述感知器（感知准则函数）算法的基本思想，并给出一种感知器学习算法。

5. （共 12 分）

（1）（4 分）请从混合高斯密度函数估计的角度，简述 K-Means 聚类算法的原理。

（2）（8 分）现有六个二维空间中的样本：$\mathbf{x}_1 = (-6,1)^T$、$\mathbf{x}_2 = (-6,-1)^T$、$\mathbf{x}_3 = (-4,0)^T$、$\mathbf{x}_4 = (4,0)^T$、$\mathbf{x}_5 = (5,1)^T$、$\mathbf{x}_6 = (6,-1)^T$。这里，上标 T 表示向量转置。请按最小距离准则对上述六个样本进行分级聚类，并画出聚类系统树图。

6. （共 15 分）

给定 d 维空间中的 n 个样本 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset R^d$，已知它们分别属于 c 个不同的类别。现在拟利用这些样本来训练一个三层前向神经网络（即包含一个输入层，一个隐含层和一个输出层）。假定采用如下交叉熵损失函数作为该网络的目标函数：$E_{ce}(\mathbf{w}) = -\sum_{k=1}^{n}\sum_{j=1}^{c} t_j^k \ln\left(z_j^k\right)$，

这里，t_j^k 表示样本 $\mathbf{x}_k$ 在输出层第 j 个结点的期望输出值（即该值已知，由样本 $\mathbf{x}_k$ 的已知类别标签来决定），z_j^k 表示样本 $\mathbf{x}_k$ 在输出层第 j 个结点的实际输出值（即通过网络计算所得的输出值），$\mathbf{w}$ 同时记录网络输入层至隐含层的各个权重 $\{w_{ih}\}$ 以及隐含层至输出层的各个权重 $\{w_{hj}\}$。请结合上述三层前向神经网络，推导误差反向传播算法，并写出具体的推导过程。

7. （12 分）

（1）（4 分）简述 PCA（主成份分析）的主要思想及其求解过程；

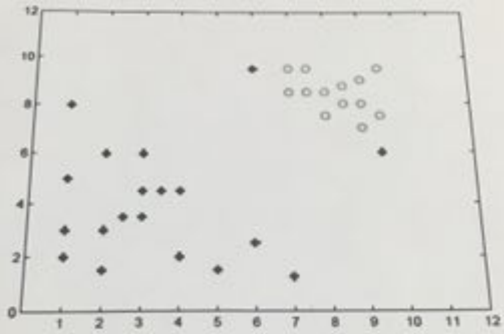
（2）（4 分）比较 PCA、CCA、LDA、ICA 的区别和适用场景；

（3）（4 分）解释 LDA（线性判别分析）所基于的数据分布假设，并阐述其不足之处。

8. （15 分）解答下面关于支持向量机（SVM）的问题。

现有一批训练数据（**有噪声**），其样本分布如图所示。现在，拟基于这些数据训练一个 SVM 分类器（二分类），以便于对测试数据进行分类。假设判别函数使用**二阶多项式核函数**。根据 SVM 原理，软间隔惩罚参数 C 会影响决策边界的位置。在下列各小题中，请定性地画出分类决策边界，**并用一两句话说明产生此边界的理由**。

（1）（3 分）当参数 C 取值特别大时（比如 $C \rightarrow \infty$），（**在答题纸上**）画出相应的分类决策边界，并说明理由：（注：按下图示所示，**先在答题纸上画出样本分布的图**）



（2）（3 分）当参数 C 取值特别小时（比如 $C \approx 0$），（**在答题纸上**）画出相应的分类决策边界，并说明理由：（注：如图所示，**先在答题纸上画出样本分布的图**）

（3）（3 分）对于（1）和（2）中的两种情形，你认为哪一种会在测试数据上表现出较好的性能，并给出相应的解释。

（4）（6 分）写出 Soft-margin SVM 的原问题及其对偶问题，并阐述核方法（kernel method）的基本思想是如何将线性模型转化为非线性模型的。