

姓名 钱之桐

学号 201828014629002

成绩

1. (共 12 分) 本题有两小题:

(1) (6 分) 对一个 c 类分类问题, 假设各类先验概率为 $P(\omega_i), i=1, \dots, c$, 条件概率密度为 $P(\mathbf{x}|\omega_i), i=1, \dots, c$ (这里 \mathbf{x} 表示特征向量), 将第 j 类模式判为第 i 类的损失为 λ_{ij} 。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) (6 分) 在 2 维特征空间, 两个类别分别有 4 个样本: $X_1=\{(3,4)^T, (3,8)^T, (2,6)^T, (4,6)^T\}$, $X_2=\{(3,0)^T, (3,-4)^T, (1,-2)^T, (5,-2)^T\}$, 假设两个类别的概率密度都为高斯分布 (正态分布) $N(\mu, \Sigma)$, 请写出两个类别的最大似然估计参数值 (μ, Σ) 。进一步, 假设两个类别先验概率相等, 请写出分类决策面的公式。

2. (共 11 分) 表示模式的特征向量 $\mathbf{x} \in R^d$, 对一个 c 类分类问题, 每一类条件概率密度为高斯分布。

(1) (5 分) 写出最小错误率决策的判别函数, 并说明在什么条件下判别函数为线性判别函数;

(2) (6 分) 当 $c=2$, 写出高斯密度条件下线性判别的决策面函数, 说明类先验概率如何影响决策面的位置, 并说明在什么情况下决策面与两个类中心差向量 $\mu_1 - \mu_2$ 垂直 (举例说明两种情况即可)。

3. (12 分) 特征空间中概率密度的非参数估计近似为 $p(\mathbf{x}) = \frac{k/n}{V}$, 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) (4 分) 说明 Parzen 窗估计和 k -近邻 (k -NN) 估计的区别;

(2) (4 分) 给定 2 维空间三个样本点 $\{(0,0)^T, (1,1)^T, (2,0)^T\}$, 请写出概率密度函数 $p(\mathbf{x})$ 的最近邻 (1-NN) 估计密度公式 (这种情况下 V 为圆形面积);

(3) (4 分) 对于 c 个类别, 基于 k -NN 概率密度估计进行贝叶斯分类, 写出各个类别的后验概率 $p(\omega_i|\mathbf{x})$ 并证明之。

4. (共 15 分)

(1) (5 分) 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;

(2) (5 分) 简述谱聚类算法的基本思想, 并指出可能影响谱聚类性能的因素;

(3) (5 分) 针对两类分类问题简述 Adaboost 算法的基本计算过程。

5. (共 10 分)

现有六个四维空间中的样本: $\mathbf{x}_1 = (0, 3, 1, 2)^T$, $\mathbf{x}_2 = (1, 3, 0, 1)^T$, $\mathbf{x}_3 = (3, 3, 0, 0)^T$, $\mathbf{x}_4 = (1, 1, 0, 2)^T$, $\mathbf{x}_5 = (3, 2, 1, 2)^T$, $\mathbf{x}_6 = (4, 1, 1, 1)^T$ 。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 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset R^d$, 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)。假定采用如下平方损失函数作为该网络的目标函数: $E(\mathbf{w}) = \sum_{k=1}^n \sum_{j=1}^c (t_j^k - z_j^k)^2$, 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} 记录所有待学习的网络参数, 包含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, 分别写出权重 w_{ih} 和权重 w_{hj} 的更新公式, 并简明扼要地给出其推导过程。

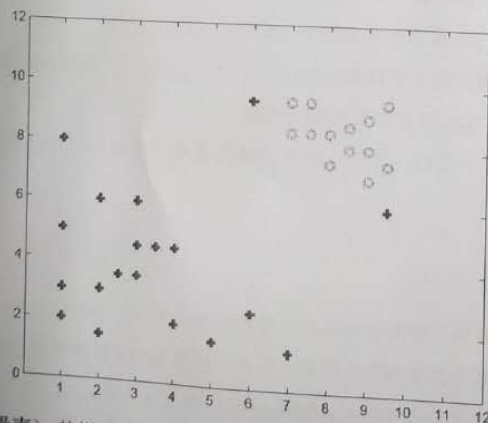
7. 数据降维 (共 8 分)

- (1) (4 分) 简述并比较 PCA、CCA、LDA、ICA 的区别和适用场景;
- (2) (4 分) 详细阐述一种实现非线性数据降维的方式。

8. 决策树 (共 8 分)

- (1) (4 分) 描述 ID3、C4.5、CART 三种决策树方法的区别;
- (2) (4 分) 阐述随机森林 (Random Forests) 的核心思想。

9. 支撑向量机 (共 9 分)



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)。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。

(1) (3 分) 当参数 C 取值特别大时 (比如 $C \rightarrow \infty$) 以及当参数 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。(注：先在答题纸上画出样本分布的图)。当 C 在什么情况下会在测试数据上表现出较好的性能，并给出相应的解释；

(2) (3 分) 对于二分类且线性可分的数据，SVM 的优化目标是最小化什么？是如何从最大化 Margin 的角度推导过来的？

(3) (3 分) 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。