

中国科学院自动化研究所

2018 年招收攻读博士学位研究生入学统一考试试卷

科目名称：模式识别

考生须知：

1. 本试卷满分为 100 分，全部考试时间总计 180 分钟。
2. 所有答案必须写在答题纸上，写在试题纸上或草稿纸上一律无效。

- 
1. (10 分). 对于  $c$  类分类问题，假定各类条件概率密度函数均为多元正态分布  $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ ,  $i=1,2,\dots,c$ ，其中， $\mathbf{x} \in R^d$  表示数据的特征向量， $\boldsymbol{\mu}_i$  和  $\boldsymbol{\Sigma}_i$  分别为第  $i$  类的均值向量和协方差矩阵， $\omega_i$  代表第  $i$  个类别。在最小错误率贝叶斯决策的框架下，请完成如下任务：

- (1)、写出一种判别函数；(6 分)
- (2)、指出在什么情况下可以获得线性判别函数，并给出具体的推导过程。(4 分)

2. (10 分). 现有一维空间中的四个样本点  $\{-2, 0, 4, 8\}$ ，请完成如下任务：

- (1)、假定窗口函数为  $\sigma=0.5$  的高斯函数，写出 Parzen 窗概率密度估计函数  $p(x)$ ；(5 分)
- (2)、写出概率密度函数  $p(x)$  的最近邻 (1-NN) 估计，画出概率密度函数曲线图。(5 分)

3. (12 分). C 均值聚类可视为是基于“误差平方和最小准则”的聚类方法，误差平方和定义为：

$$J = \sum_{i=1}^c J_i, \quad J_i = \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2, \quad \mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

其中， $c$  是聚类个数， $D_i$  表示第  $i$  个聚类， $\mathbf{m}_i$  表示属于第  $i$  个聚类的所有样本的均值， $n_i$  表示属于第  $i$  个聚类的样本个数， $J_i$  表示第  $i$  个聚类的误差平方和， $J$  是总的误差平方和。

- (1)、写出一种 C 均值聚类算法的计算步骤；(7 分)
- (2)、在某次迭代时，假定原来属于  $D_i$  个聚类中的样本  $\hat{\mathbf{x}}$  被分配至  $D_j$  个聚类，试证明这种分配会导致  $J_j$  增加为：  $J_j^* = J_j + \frac{n_j \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2}{n_j + 1}$ 。(5 分)

4. (10 分). 关于 AdaBoost 方法。假定有  $n$  个  $d$  维空间中的训练样本  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset R^d$ ，进一步假定这些样本属于两个不同的类别。请完成如下任务：

- (1)、基于这些样本，请简述采用 AdaBoost 方法学习一个两类分类器的计算步骤；(6 分)
- (2)、请解释为什么 AdaBoost 经常可以在训练误差为零时继续训练还可以带来测试误差继续下降。(4 分)

(未完待续)

5. (20 分). 在一个模式识别问题中, 有下列 8 个样本, 每个样本为一个二维特征向量:  $(-4, 1)^T$ ,  $(-2, 1)^T$ ,  $(-4, -1)^T$ ,  $(-2, -1)^T$ ,  $(4, 1)^T$ ,  $(2, 1)^T$ ,  $(4, -1)^T$ ,  $(2, -1)^T$ 。其中, 括号内的第一个数据表示该样本点的第一个特征, 括号内的第二个数据表示该样本点的第二个特征, 上标  $T$  表示向量转置。
- (1)、如果不知道这 8 个样本的类别标签, 请采用 K-L 变换, 计算其特征值和特征向量。  
(10 分)
- (2)、对上述 8 个样本, 假设前 4 个样本属于第一类, 后 4 个样本属于第二类, 现在要求只使用一个特征达到这样的分类目的。请采用  $J_s = \frac{|\mathbf{S}_b + \mathbf{S}_w|}{|\mathbf{S}_w|}$  判据对特征进行选择, 给出计算过程和结果。(10 分)
6. (16 分). 已知正类样本点  $\mathbf{x}_1=(1, 1)^T$ ,  $\mathbf{x}_2=(1, 0)^T$ ,  $\mathbf{x}_3=(3, 3)^T$ , 负类样本点  $\mathbf{x}_4=(4, 3)^T$ ,  $\mathbf{x}_5=(4, 4)^T$ ,  $\mathbf{x}_6=(5, 3)^T$ 。请完成如下任务:
- (1)、写出线性支持向量机需要求解的原问题和对偶问题 (不要求进行求解); (10 分)
- (2)、当软间隔惩罚参数  $C$  取值很大时, 定性地画出所得到的分类决策面, 并解释原因;  
(提示: 先将 6 个样本点在答题纸上画出, 然后画出分类决策面) (3 分)
- (3)、当软间隔惩罚参数  $C$  取值很小时, 定性地画出所得到的分类决策面, 并解释原因。  
(提示: 先将 6 个样本点在答题纸上画出, 然后画出分类决策面) (3 分)
7. (12 分). 针对多层前馈神经网络, 请回答如下问题:
- (1)、结合三层网络分别给出有关“隐含层至输出层”、“输入层至隐含层”的权重更新公式;  
(4 分)
- (2)、给出采用误差反向传播算法 (即 BP 算法) 训练一个多层前馈神经网络的计算步骤;  
(4 分)
- (3)、在采用 BP 算法对网络进行训练时, 在训练过程中可能产生麻痹现象。解释为什么会出现这种现象。 (4 分)
8. (10 分). 关于深度学习, 请回答如下问题:
- (1)、请对“深度学习”下一个定义; (4 分)
- (2)、请列举出三种不同的深层神经网络模型 (给出其名称即可), 并指出它们分别适用于何种模式识别任务。(6 分)