《模式识别与机器学习》期末考试试题（A卷）

姓名：_____　　班级：_____　　学号：_____

任课教师：汪小我、张学工

考试时间：2022年6月7日10:00-12:00

请检查试卷页数：共 8 页

一、 不定项选择题（每题 4 分，共 24 分，每小题漏选得 2 分，错选或不选不得分）

1. 以下深度学习常用技巧中，可以用来处理网络过拟合的是：（ ）

A. 添加 L1/L2 正则项

B. 添加 Dropout 层

C. 数据增强

D. 将部分测试集用于训练

2. 在方差已知的条件下，采用贝叶斯估计方法估计正态分布的均值，先验分布取均值μ_0、方差σ_0^2的正态分布，下列说法**正确**的是：（ ）

A. 选取不同的损失函数会影响估计结果

B. σ_0^2越大，贝叶斯估计的结果越接近μ_0

C. 样本数越多，贝叶斯估计的结果越接近样本均值

D. 若有新样本出现，需要和之前所有样本一起从头计算参数的后验概率

3. 图 1.1 展示了 adaboost 算法训练过程中，第一个分类器对四个样本的输出结果，其中○表示正样本，×表示负样本。设样本初始权重相等，下列说法**正确**的是：（ ）

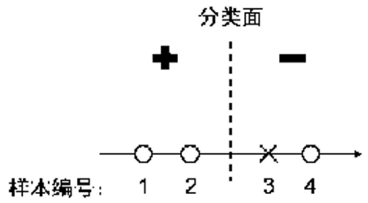

图 1.1 adaboost 分类结果示意图

A. 图中的四个样本线性可分

B. 本轮结束后，将增加 3 号样本的权重

C. 本轮结束后，样本的权重可能更新为 1/8、1/8、1/8、5/8

D. 继续迭代，adaboost 算法的训练正确率最终能达到 100%

4. 回顾一致聚类中的一致性矩阵$\mathcal{M}_{N\times N}$,，其元素为 0~1 之间的实数，反映了两个样本被聚为同一类的频率。据此，可定义 CDF 函数如下：

$$CDF^{(k)}(t)=\frac{\sum_{i<j}I\{\mathcal{M}^k(i,j)\le t\}}{N(N-1)/2}$$

其中N为样本数量，k为聚类数，$I(\cdot)$为指示函数。不同聚类数k下的 CDF 函数曲线如图 1.2 所示，以下说法**正确**的是：（ ）

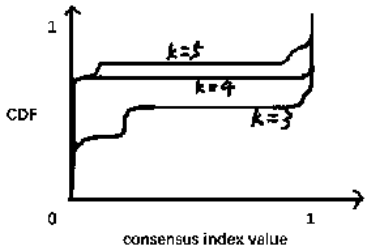

图 1.2 CDF 函数曲线图

A. 在根据 CDF 曲线选择k时，应当选择曲线下面积（AUC）最大的k值

B. 在根据 CDF 曲线选择k时，应当选择曲线下面积（AUC）最小的k值

C. 根据图中的结果，应当选择$k=5$作为聚类类别数

D. 根据图中的结果，应当选择$k=4$作为聚类类别数

5. 下列关于支持向量机（SVM）的说法中，**正确**的是（ ）

A. 线性核函数的 SVM 一定会得到线性分类面

B. 非线性核函数的 SVM 一定会得到非线性分类面

C. 如果数据集线性可分，则线性 SVM 中支持向量到分类面的距离是$\frac{1}{\|w\|}$

D. 在线性 SVM 中引入松弛变量，可以得到非线性分类面

6. 在隐马尔可夫(HMM)模型中，若观测值不仅与当前时刻的隐状态直接相关，还与上一时刻的隐状态直接相关，如图 1.3 所示，对于该新模型，考虑节点为 0/1 取值，下列说法正确的是：（ ）

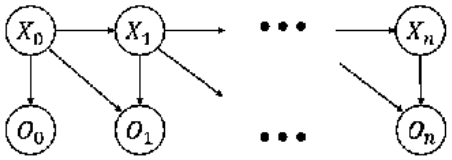

图 1.3 HMM 修改后的新模型示意图

A. 该模型中存在环，所以不是贝叶斯网络

B. 该模型中存在环，但仍是贝叶斯网络

C. 若观测序列的隐状态已知，各时刻的观测变量间相互独立

D. 若观测序列的隐状态已知，可使用最大似然估计模型参数

二、 填空题（20 分）

1. 判断以下陈述是否正确（每题 1 分，共 4 分）

（1） 如果数据线性不可分，则感知器不会收敛；如果数据线性可分，则感知器有唯一解。（ ）

（2） 特征选择时高度相关的特征是完全冗余的，没有必要保留。（ ）

（3） L1 范数与 L2 范数正则化都是保凸的，且使用 L1 范数更容易得到稀疏解。（ ）

（4） 二分类问题中，随机分类的 AUROC 的期望为 0。（ ）

2. 用随机梯度下降法**训练**网络时，从以下选项中**选择**合适的步骤并排序：（4 分）

① 将抽取的数据输入网络，得到预测结果
② 将预测结果和标签比较，计算损失函数值
③ 迭代至满足终止条件，结束训练
④ 按 Batch 依次从训练集中抽取一定数量的数据
⑤ 更新网络参数
⑥ 当前 epoch 结束后，计算网络在验证集上的损失函数值
⑦ 按 Batch 依次从验证集中抽取一定数量的数据
⑧ 将数据集划分为训练集、验证集、测试集
⑨ 计算损失函数对网络各参数的导数
⑩ 当前 epoch 结束后，计算网络在测试集上的损失函数值

3. 现欲采用k-近邻算法对图 2.1 所示的○、×两类样本分类。假设$k=1$，请在图 2.1 中画出最近邻算法的分类面（2 分）。假设$k=3$，采用留一法做交叉验证的正确率为_____（2 分）。

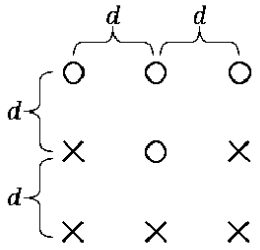

图 2.1 样本分布图

4. 某种疾病检测筛查的混淆矩阵如下，请根据表中数据，计算该次检测的正确率（accuracy）_____，灵敏度（sensitivity）_____，特异度（specificity）_____。若该疾病在人群中的发病率为 1%，某人连续两次**独立**检测呈阳性，则他患有疾病的概率为_____（每空 2 分，共 8 分）。

表 2.1 混淆矩阵

真实标签	检测标签	
	阳性（P）	阴性（N）
阳性（P）	90	10
阴性（N）	10	990

可能用到的公式：

正确率（accuracy）：$ACC = \frac{TP+TN}{TP+TN+FP+FN}$

灵敏度（sensitivity）：$TPR = \frac{TP}{TP+FN} = 1 - FNR$

特异度（specificity）：$TNR = \frac{TN}{TN+FP} = 1 - FPR$

三、 Fisher 线性判别（8 分）

已知有两类样本ω_1和ω_2：

$$\omega_1 = \{(1,1),(1,2),(1,3),(1,4),(1,5)\}$$
$$\omega_2 = \{(2,5),(3,5),(4,5),(5,5),(6,5)\}$$

请使用 Fisher 线性判别计算出上述数据的最优投影方向（6 分）。若有新样本点$(3,3)$，请选择合适的分类阈值对新样本分类，说明你的分类过程（2 分）。

可能用到的公式：

类均值向量：$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{x\in C_i}\boldsymbol{x}$，$i = 1,2$（$N_i$是第$i$类样本集合$C_i$的样本数）

类内离散度矩阵：$\boldsymbol{S}_i = \sum_{x\in C_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T, i = 1,2$

总类内离散度矩阵：$\boldsymbol{S}_w = \boldsymbol{S}_1 + \boldsymbol{S}_2$

类间离散度矩阵：$\boldsymbol{S}_b = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$

Fisher 准则的目标函数：$max_{\boldsymbol{w}}J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$，其解为$\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$

四、 主成分分析（8 分）

对于题目“Fisher 线性判别”中的样本点，采用主成分分析（PCA）的方法将其降到一维，计算投影矩阵（6 分）。同样，对新样本点$(3,3)$降维后采用最近邻算法分类，说明你的分类过程（2 分）。

表 5.1 风险矩阵

	ω_1	ω_2
α_1	0	λ_{12}
α_2	λ_{21}	0

五、 正态分布的贝叶斯决策（8 分）

现有两类样本分别来自**方差相等**的两个一维正态分布：

$$p(x|\omega_1) = \mathcal{N}(0,\sigma^2), \qquad p(x|\omega_2) = \mathcal{N}(1,\sigma^2)$$

设两类样本的先验概率分别为$P(\omega_1)$和$P(\omega_2)$，风险矩阵如表 5.1 所示。其中，ω_i表示样本实际为第i类，α_i表示决策为第i类。试求最小风险准则下贝叶斯决策的分类阈值t。

六、 决策树（8 分）

心脏病是一类常见的循环系统疾病，为了预测病人是否患有心脏病，表 6.1 统计了 6 名志愿者的性别（G）、是否胸痛（P）、是否吸烟（S）、是否经常锻炼（E）、心率（R）及其是否患有心脏病（H）的信息。

1. 仅使用 G、P、S、E 四维特征预测 H，请使用信息熵作为信息度量建立决策树，写出树的计算生成过程（4 分），并画出最终的树结构（2 分）。
2. 若想利用心率（R）信息优化预测性能，有哪几种决策树算法可以采用，请简要说明处理方法（2 分）。

表 6.1 心脏病信息表

编号	性别 G	胸痛 P	吸烟 S	锻炼 E	心率 R	心脏病 H
1	男	是	否	是	57	是
2	男	是	是	否	92	是
3	女	否	是	否	103	是
4	男	否	否	是	77	否
5	女	是	是	是	85	是
6	男	否	是	是	63	否

表 6.2 $log_2 x$对数表

x	1	1/2	1/3	2/3	1/4	3/4
$log_2 x$	0	-1	-1.58	-0.58	-2	-0.42

七、 k-均值聚类（8 分）

设有n个样本点$\{x_1, x_2, \ldots x_n\}$，采用k-均值算法将其聚为k类，记k类样本中心分别为$\{m_1, m_2, \ldots m_k\}$，请回答下列问题：

1. 标准的k-均值算法通过计算当前聚类中所有样本点的均值来更新聚类中心，该过程的伪代码如下：

输入： 样本集$D=\{x_j, j=1,2,\dots,n\}$，聚类数k

初始化聚类中心$\{m_1, m_2, \dots m_k\}$

while not convergence:

令聚类集合$\Gamma_i=\emptyset, i=1,2,\dots,k$

for $j=1,2,\dots,n$ **do**

计算样本x_j与各聚类中心$m_i, i=1,2,\dots,k$的距离$d_{ji}=\|x_j-m_i\|_2$

确定样本x_j的聚类标记$\lambda_j=\arg\min_i d_{ji}$

将样本x_j划分至对应聚类$\Gamma_{\lambda_j}\leftarrow\Gamma_{\lambda_j}\cup\{x_j\}$

end for

for $i=1,2,\dots,k$ **do**

更新聚类中心$m_i\leftarrow\frac{1}{|\Gamma_i|}\sum_{x\in\Gamma_i}x$，$|\Gamma_i|$表示聚类$\Gamma_i$中样本点的数量

end for

输出： 聚类划分$\{\Gamma_i, i=1,2,\dots,k\}$

现有 10 个一维样本点$\{-5,-4,-2.1,-1.7,1.5,\ 3,\ 5.2,\ 7,\ 8.1,10\}$，设$k=2$，初始聚类中心分别设为$\{-3,\ 3\}$，请计算运行**一步**$k$-均值算法后的聚类中心（2 分）。

2. 若聚类损失函数定义如下：

$$L=\sum_{i=1}^{k}\sum_{x\in\Gamma_i}\|x-m_i\|_2^2$$

请以第i类为例，证明m_i的更新过程等价于学习率为$\frac{1}{|\Gamma_i|}$的梯度下降过程（6 分）。

八、 神经网络（16 分）

小睿是某视频网站的审核员。他的工作内容是，观看用户上传的视频，将有违法违规内容的视频标记为“不合格”，其他的视频的则为“合格”。网站会将审核标记为合格的视频正式上线。

随着网站规模的扩大，审核部门的工作负担也越来越重，时常需要加班才能完成任务。热爱生活的小睿讨厌加班，于是他设计了一套自动视频审核系统，辅助自己完成审核。以下是他设计的系统（如图 8.1 所示）：

输入： 维度为$[T,C,H,W]$的待审核视频，即视频总共包含T帧图像，每张图像有C个通道，图像的高度和宽度分别为H,W。

输出： $[0,1]$上的实数，表示待审核视频合格的概率。

1. 这是一个二分类问题，训练网络时常使用 BCE 损失：

$$\mathcal{L}(y_{\text{gt}}, y_{\text{pred}}) = -[y_{\text{gt}} \log(y_{\text{pred}}) + (1 - y_{\text{gt}}) \log(1 - y_{\text{pred}})]$$

请写出$\frac{\partial \mathcal{L}}{\partial y_{\text{pred}}}$的表达式（1 分）。

2. 小睿的数据集由大量已上线的视频和少量刚完成审核的视频组成。自然地，这样的数据集内合格视频的数量远多于不合格视频。实际审核中，小睿并不希望漏检太多的不合格视频，即需要分类系统对不合格视频的召回率尽可能高。请问在这种情况下，可采取何种方法提升对不合格视频的召回率？（可以考虑的方面包括但不限于损失函数、训练过程、部署推理部分的修改，**答一条即可**，2 分）

以下是具体的网络结构示意图和思路说明

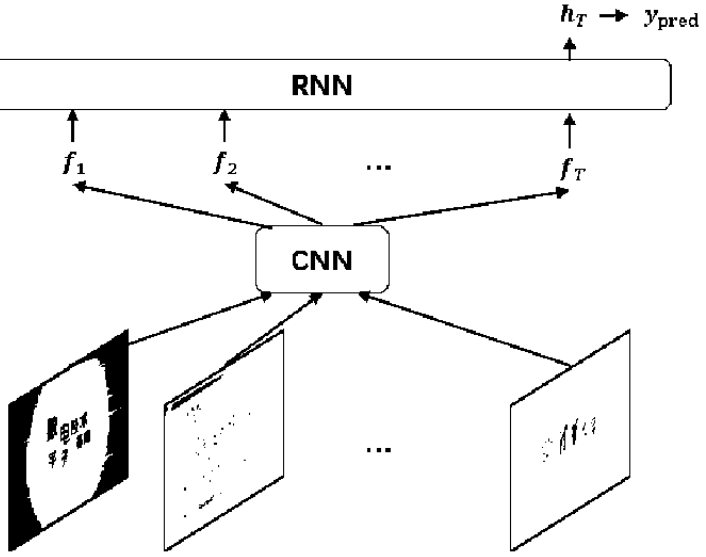

图 8.1 视频审核系统网络示意图

具体思路：对于输入视频$V \in \mathbb{R}^{T \times C \times H \times W}$，经历如下步骤得到预测结果：

- 拆帧。将视频按照播放顺序拆分为图片组$\{I_1, I_2, \ldots, I_T\}$，其中$I_i \in \mathbb{R}^{C \times H \times W}$
- 将图片组送入**同一**卷积网络，得到图片深度特征$\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$，其中$\boldsymbol{f}_i \in \mathbb{R}^n$
- 将图片深度特征按序输入循环神经网络，由循环神经网络聚合，得到**最后一个时间步**的隐层状态$\boldsymbol{h}_T \in \mathbb{R}^{\text{m}}$，作为整个视频的深度特征
- 将隐层状态经过$\mathbb{R}^{\text{m}} \rightarrow [0,1]$的线性分类器，即可预测视频是否合格

3. 卷积网络部分的具体设计如下，设所有卷积层的$\text{padding} = 0$，$\text{stride} = 1$，池化层的$\text{padding} = 0$，$\text{stride} = 2$。请填写表 8.1 中相关模块的参数个数和输出维度（6 分）。

表 8.1 神经网络各层参数表

层名称	输出维度	权值参数数量	偏置参数数量
输入	3*1920*1080		
卷积层(9*9,32 个)			
池化层(2*2)			
卷积层(5*5,64 个)			
池化层(3*3)			
卷积层(5*5,64 个)			
自适应池化层	64*1*1		

4. 循环神经网络部分的具体设计如下：

$$\boldsymbol{h}_t = \tanh(\boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{f}_t), \qquad t = 1,2,\dots,T, \boldsymbol{h}_0 = \boldsymbol{0}$$

$$y_{\text{pred}} = \text{sigmoid}(\boldsymbol{V}\boldsymbol{h}_T + d)$$

按照题设约定，输入、输出、网络参数的维度如表 8.2：

表 8.2 网络输入、输出、参数维度表

变量名	维度
$\mathcal{L}$	$\mathbb{R}$
y_{pred}	$\mathbb{R}$
y_{gt}	$\mathbb{R}$
$\boldsymbol{h}_t(t = 0,1,\dots,T)$	$\mathbb{R}^m$
$\boldsymbol{f}_t(t = 1,2,\dots,T)$	$\mathbb{R}^n$
$\boldsymbol{V}$	$\mathbb{R}^{1\times m}$
$\boldsymbol{W}$	$\mathbb{R}^{m\times m,}$
$\boldsymbol{U}$	$\mathbb{R}^{m\times n,}$
$\boldsymbol{b}$	$\mathbb{R}^m$
d	$\mathbb{R}$

用一个标签为y_{gt}的样本训练网络，batch size = 1，梯度下降的学习率为η，写出网络参数$d,\boldsymbol{V},\boldsymbol{W}$的更新过程（7 分，$d,\boldsymbol{V},\boldsymbol{W}$分别占 1、2、4 分）。

可能用到的公式：

tanh 函数求导公式：$\tanh'(x) = 1 - \tanh^2(x)$

sigmoid 函数求导公式：$\text{sigmoid}'(x) = \text{sigmoid}(x)\cdot\big(1 - \text{sigmoid}(x)\big)$