中国科学院大学

考试日期：2020.01.11

试 题 专 用 纸

课程编号：721004H

课程名称：模式识别与机器学习

任课教师：黄庆明等

姓名_____ 学号_____ 成绩_____

一、（16 分）选择题。（每个选项 2 分，请将答案写在答题纸上）

1. 基于二次准则函数的 H-K 算法较之于感知器算法的优点是哪个？

A. 计算量小

B. 可以判别问题是否线性可分

C. 其解完全适用于非线性可分的情况

2. 在逻辑回归中，如果正则项取 L1 正则，会产生什么效果？

A. 可以做特征选择，一定程度上防止过拟合

B. 能加快计算速度

C. 在训练数据上获得更准确的结果

3. 如果模型的偏差较高，我们如何降低偏差？

A. 在特征空间中减少特征

B. 在特征空间中增加特征

C. 增加数据点

4. 假设采用正态分布模式的贝叶斯分类器完成一个两类分类任务，则下列说法正确的是哪个。

A. 假设两类的协方差矩阵均为对角矩阵，则判别界面为超平面。

B. 假设两类的协方差矩阵相等，则判别界面为超平面。

C. 不管两类的协方差矩阵为何种形式，判别界面均为超平面。

5. 下列方法中，哪种方法不能用于选择 PCA 降维（K-L 变换）中主成分的数目K？

A. 训练集上残差平方和随K发生剧烈变化的地方（肘部法）

B. 通过监督学习中验证集上的性能选择K

C. 训练集上残差平方和最小的 K

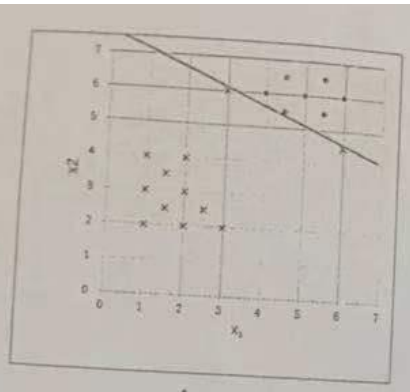
6. 考虑某个具体问题时，你可能只有少量数据来解决这个问题。不过幸运的是你有一个针对类似问题已经预先训练好的神经网络，请问可以用下面哪种方法来利用这个预先训练好的网络？

A. 把除了最后一层外所有的层都冻住，重新训练最后一层

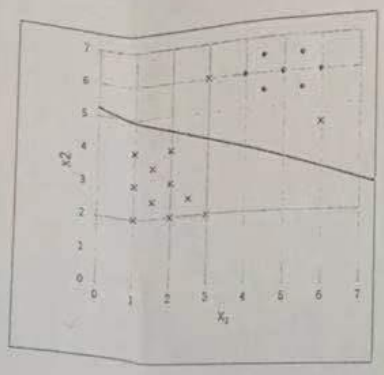
B. 对新数据重新训练整个模型

C. 只对最后几层进行调参(fine tune)

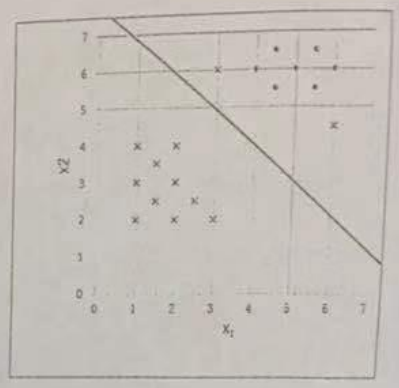
7. 如下图所示，假设该数据集中包含一些线性可分的数据点。训练 Soft margin SVM 分类器，其松弛项的系数为C。请问当$C \to 0$时，分类边界为下图中的哪个？



A

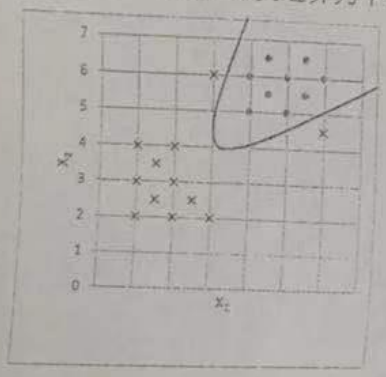


B

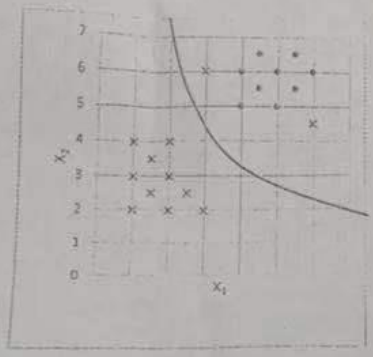


C

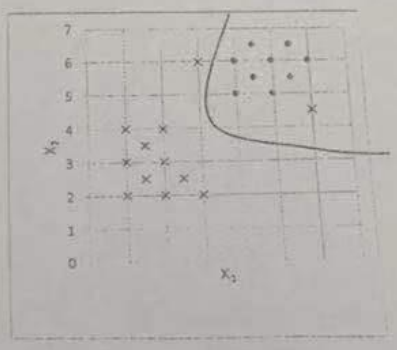
8. 如下图所示，假设该数据集中包含线性不可分的数据点。采用二次核函数训练 Soft margin SVM 分类器，请问当$C \to \infty$时，分类边界为下图中的哪个？



A



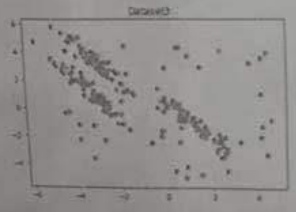
B



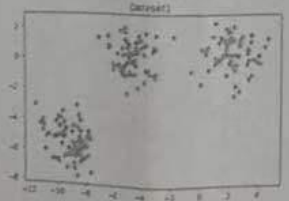
C

二、（6 分）请列举半监督学习对数据样本的三种基本假设。

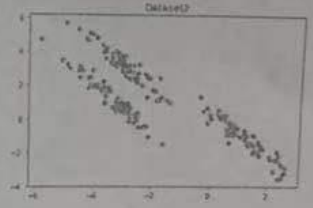
三、（8 分）针对下图所示的三种数据分布，从 K 均值、GMM 和 DBSCAN 中分别选择最合适的聚类算法，并简述理由。



(a)



(b)



(c)

四、（12 分）对于具有类别标签的数据，采用 K-L 变换和 Fisher 线性判别分析两种方法对数据降维。

（1）简述这两种数据降维方法的基本过程。（8 分）

（2）这两种方法中哪种方法对分类更有效？并简述原因。（4 分）

五、（10 分）逻辑回归

（1）简述逻辑回归算法的原理。（4 分）

（2）如果使用逻辑回归算法做二分类问题得到如下结果，分别应该采取什么措施以取得更好的结果？并说明理由。（6 分）

（a）训练集的分类准确率 85%，验证集的分类准确率 80%，测试集的分类准确率 75%；

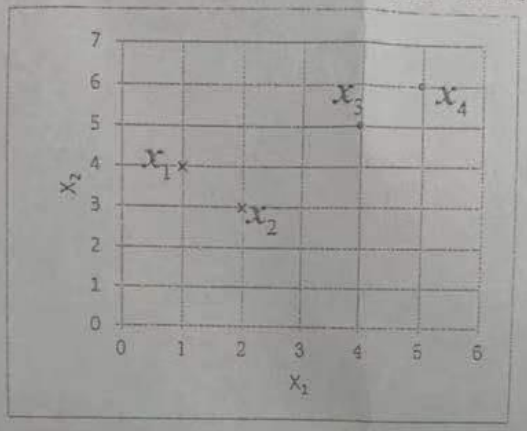
（b）训练集的分类准确率 99%，验证集的分类准确率 80%，测试集的分类准确率 78%；

六、（10 分）解释 AdaBoost 算法的基本思想和工作原理，并给出 AdaBoost 算法的伪代码。

七、（10 分）从特征提取的角度，分析深度卷积神经网络与传统特征提取方法（例如 Gabor 小波滤波器）的异同，并给出深度学习优于传统方法的原因。

八、（8 分）硬间隔支持向量机（Hard margin SVM）

如下图所示，一个数据集包含来自 2 个类别的 4 个数据点，在此集合上训练一个线性 Hard margin SVM 分类器。请写出 SVM 的形式化模型，并计算出该分类器的权重向量w和偏差b，给出该分类器的支持向量。



九、（10 分）拟利用贝叶斯判别方法检测 SNS 社区中不真实账号。设$Y = 0$表示真实账号，$Y = 1$表示不真实账号。每个用户有三个属性：X_1表示日志数量/注册天数，X_2表示好友数量/注册天数，X_3表示是否使用真实头像。已知$P(Y = 0) = 0.89$，$P(X_3 = 0|Y = 0) = 0.2$，$P(X_3 = 0|Y = 1) = 0.9$，且给定Y的情况下X_1、X_2的分布如下：

$P(X_1\|Y)$	$X_1 \le 0.05$	$0.05 < X_1 \le 0.2$	$X_1 \ge 0.2$
$Y = 1$	0.8	0.1	0.1
$Y = 0$	0.3	0.5	0.2
$P(X_2\|Y)$	$X_2 \le 0.1$	$0.1 < X_2 \le 0.8$	$X_2 \ge 0.8$
$Y = 1$	0.7	0.2	0.1
$Y = 0$	0.1	0.7	0.2

若一个账号使用非真实头像，日志数量与注册天数的比率为 0.1，好友数与注册天数的比率为 0.2，判断该账号是不是虚假账号。

十、（10 分）现装有红色球和白色球的两个盒子，盒子 1 中红球的比例为p，盒子 2 中红球的比例为q。我们以概率π选择盒子 1,概率$1-\pi$选择盒子 2，然后从盒子中有放回地取出一个小球，独立地重复进行 4 次试验，观测结果为：红，红，白，红。

假定模型的参数初始值为$\pi^{(0)} = 0.4$，$p^{(0)} = 0.4$，$q^{(0)} = 0.5$，请写出 EM 算法迭代一次后p和q的值。（计算结果保留两位小数）