

中南大学考试试卷 (A卷)

2020 -- 2021 学 年 下 学 期

时间 100 分钟

2021 年 7 月 2 日

机器学习 课程 48 学时 3 学分 考试形式: 开 卷

专业年级: 信安 18 级

总分 100 分, 占总评成绩 50%

注: 此页不作答题纸, 请将答案写在答题纸上

一、选择题 (本题 15 分, 每小题 3 分)

1. 已知对一组观察值 (x_i, y_i) , $i = 1, 2, \dots, N$, 作出散点图后发现数据具有线性相关关系, 若假设 $h_\theta(x) = \theta_0 + \theta_1 x$, 已知其中一个数据对为 $(0, 1)$, 且 $\frac{1}{N} \sum_{i=1}^N x_i = 5$, $\frac{1}{N} \sum_{i=1}^N y_i = 3.5$, 则线性回归方程为 ()
A、 $h_\theta(x) = 1 + 0.5x$ B、 $h_\theta(x) = 1 + 0.7x$
C、 $h_\theta(x) = 1 - 0.5x$ D、 $h_\theta(x) = 1 - 0.7x$
2. 关于主元分析的应用, 下列描述**错误**的是 ()
A、数据压缩 B、数据降维 C、数据可视化 D、减少过拟合
3. Logistic 回归的输出可以作为概率, 假定训练好的 Logistic 回归假设 $h_\theta(x)$ 对新样本 x 的输出 $h_\theta(x) = 0.2$, 意味着 ()
A、 $P(y = 1|x; \theta) = 0.2$ B、 $P(y = 1|x; \theta) = 0.8$
C、 $P(y = 0|x; \theta) = 0.2$ D、 $P(y = 0|x; \theta) = 0.1$
4. 关于采用随机森林替代决策树, 下列描述中**正确**的是 ()
A、减少偏差 B、增加模型的可解释性 C、增加偏差 D、减少过拟合
5. 关于 Soft SVM, 减小松弛因子 C 会导致 ()
A、增加过拟合 B、减少间隔 C、对训练集中的 outlier 更敏感 D、减少过拟合

二、PCA (15分)。给定2维数据如图1所示, 并给出了a, b, c, d四个投影方向, 试回答

- (1) 第一个主元投影方向是? 为什么?
- (2) 第二个主元投影方向是? 为什么?
- (3) 第三个主元投影方向是? 为什么?

三、决策树 (15分)。给定2维训练集如图2所示, 试训练CART决策树, 使得该决策树在该训练集上的错误率为0。

四、Adaboost (20分)。针对如图2所示的训练集, 试采用Adaboost算法训练一个由4个弱分类器 (decision stump) 组成的强分类器, 并给出该强分类器的训练误差。

五、神经网络 (20分)。假设有如下结构的神经网络结构 (如图3所示):

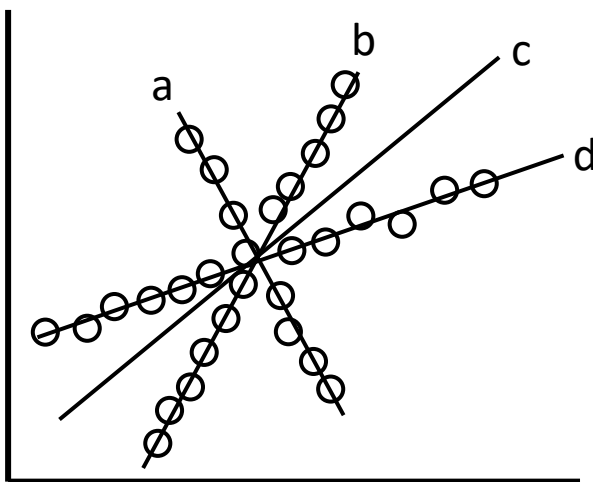


图1

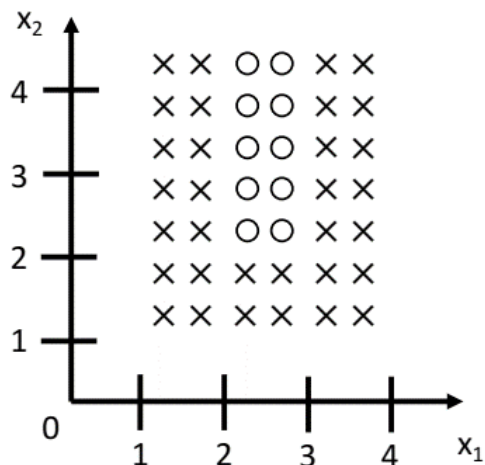


图2

其中每个神经元均采用Relu激活函数，其权重初始化为

$$\Theta_{10}^{(1)} = -0.4, \Theta_{11}^{(1)} = 0.2, \Theta_{12}^{(1)} = 0.1, \Theta_{20}^{(1)} = -0.2, \Theta_{21}^{(1)} = 0.4, \Theta_{22}^{(1)} = -0.1$$

$$\Theta_{10}^{(2)} = 0.1, \Theta_{11}^{(2)} = -0.2, \Theta_{12}^{(2)} = 0.1, \Theta_{20}^{(2)} = 0.4, \Theta_{21}^{(2)} = -0.1, \Theta_{22}^{(2)} = 0.1$$

- (1) 给定样本 $x = [1, 0]^T$ ，试给出所有隐藏层节点的输入和输出；(8分)
- (2) 若 x 对应的 ground-truth 为 $y = [0.9, 0.1]^T$ ，采用带有 Momentum 的梯度下降法对根据该样本采用反向传播算法更新参数，学习率为 0.1，Momentum $\rho = 0.9$ ，试给出第二次更新后的参数 $\Theta_{10}^{(1)}$ 的值。(12分)

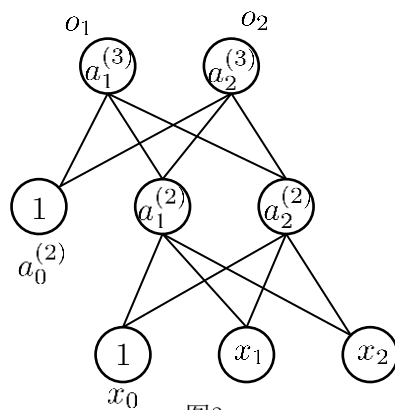


图3

六、混淆矩阵（15分）。人们利用临床和医学影像数据，采用机器学习方法构建预测疑似病人是否真正患有新冠肺炎的模型。假设采用 Logistic Regression 训练了 2 个分类器 C1 和 C2 来预测是否患有新冠肺炎，这两个分类器在在一个具有 8 个样本的测试集测试结果如下表所示其中 y_{gt} 为样本的实际类别（1 表示阳性，0 表示阴性）， y_{C1} 和 y_{C2} 分别为分类器 C1 和 C2 的输出。根据上述数据：

y_{gt}	1	0	1	0	1	0	0	1
y_{C1}	0.7	0.3	0.88	0.22	0.4	0.2	0.75	0.62
y_{C2}	0.4	0.5	0.34	0.55	0.7	0.6	0.53	0.33

- (1) 针对分类器 C1，若阈值设置为 0.5，给出对应的混淆矩阵，并计算其的召回率或敏感性(sensitivity)、特异性(specificity)；
- (2) 针对分类器 C2，阈值设置为 0.3，同样给出对应的混淆矩阵，并计算其的召回率、精度(precision)和 F1-score。