西北工业大学研究生院

学 位 研 究 生 课 程 考 试 试 题

考试科目：机器学习方法及应用　　课程编号：M08M11078
开课学期：2019.3-2019.6　　考试时间：2019.7.10
说　　明：所有答案必须写在答题册上，否则无效。　　共 3 页　第 1页

1. 选择题 （共 15 分，每小题 5 分）

（1）如果一个 SVM 模型出现欠拟合，那么下列哪种方法能解决这一问题？

A. 增大惩罚参数 C 的值

B. 减小惩罚参数 C 的值

C. 减小核系数（gamma 参数）

（2）我们知道二元分类的输出是概率值。一般设定输出概率大于或等于 0.5，则预测为正类；若输出概率小于 0.5，则预测为负类。那么，如果将阈值 0.5 提高，例如 0.6，大于或等于 0.6 的才预测为正类。则准确率（Precision）和召回率（Recall）会发生什么变化（多选）？

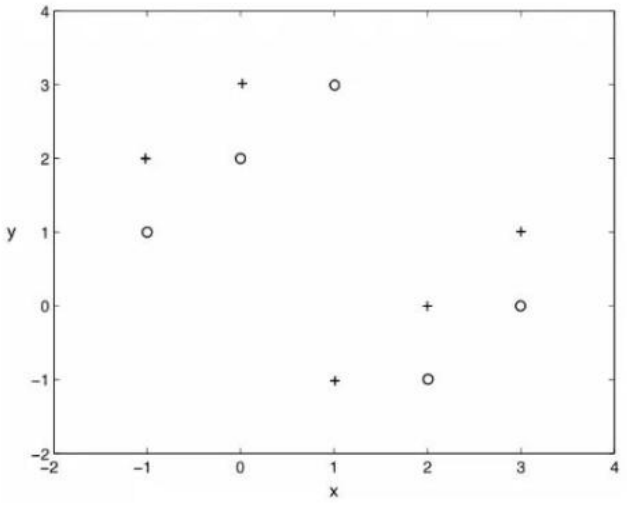
A. 准确率（Precision）增加或者不变

B. 准确率（Precision）减小

C. 召回率（Recall）减小或者不变

D. 召回率（Recall）增大

（3）假设我们使用 kNN 训练模型，其中训练数据具有较少的观测数据（下图是两个属性 x、y 和两个标记为 “+” 和 “o” 的训练数据）。现在令 k = 1，则图中的 Leave-One-Out 交叉验证错误率是多少？



A. 0%

B. 20%

C. 50%

D. 100%

（4）下列哪些算法可以用来够造神经网络（多选）？

A. kNN

B. 线性回归

C. 逻辑回归

（5）数据科学家经常使用多个算法进行预测，并将多个机器学习算法的输出（称为“集成学习”）结合起来，以获得比所有个体模型都更好的更健壮的输出。则下列说法正确的是？

A. 基本模型之间相关性高

B. 基本模型之间相关性低

C. 集成方法中，使用加权平均代替投票方法

D. 基本模型都来自于同一算法

2. 概念学习（15 分）

请看以下的正例和反例，它们描述的概念是“两个住在同一房间中的人”。每个训练样例描述一个有序对，每个人由其性别、头发颜色（black、brown 或 blonde）、身高（tall、medium 或 short）以及国籍（US、French、German、Irish、Indian、Chinese 或 Portuguese）。

＋ <<male brown tall US>, <female black short US>>

＋ <<male brown short French>, <female black short US>>

－<<female brown tall German>, <female black short Indian>>

＋<<male brown tall Irish>, <female brown short Irish>>

考虑在这些实例上定义的假设空间为：所有假设以一对 4 元组表示，其中每个值约束与 EnjoySport 中的假设表示相似，可以为：特定值、“？”或者“∅”。例如，下面的假设：

<<male ? tall ?> <female ? ? French>>

它表示了所有这样的有序对：第一个人为高个男性（国籍和发色任意），第二个为法国女性（发色和身高任意）。

(a) 根据上述提供的训练样例和假设表示手动执行候选消除算法。特别是要写出处理了每一个训练样例后变型空间的特殊和一般边界。

(b) 计算给定假设空间中有多少假设与下面的正例一致：

+ <<male black short Portuguese>, <female blonde tall Indian>>

(c) 如果学习器只有一个训练样例，如(b)中所示，现在由学习器提出查询，并由施教者给出其分类。求这一特定的查询序列，以保证学习器能收敛到单个正确的假设，而不论该假设是哪一个（假定目标概念可以使用给定的假设表示语言来描述）。求出最短的查询序列，这一序列的长度与问题(b)的答案有什么关联？

(d) 注意到这里的假设标示语言不能够表示这些实例上的所有概念（如我们可以定义出一系列的正例和反例，它们并没有相应的可描述假设）。如果要扩展这一语言，使其能够表达该实例语言上的所有概念，那么(c)的答案应该如何更改。

3. 决策树（15 分）

以下的数据集被用来学习一个决策树，这个决策树将体重（Normal 或 Underweight）、眼睛的颜色（Amber 或 Violet）和视力（2 或 3 或 4）等作为特征量来预测学生是懒惰的（L）还是勤奋的（D）。

Weight	Eye Color	Num. Eyes	Output
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

请回答以下问题，不需要写出推导过程：

(1) 条件熵 H(EyeColor|Weight = N)是多少？

(2) ID3 算法会选择哪个属性作为树根（不修剪）？

(3) 画出从这些数据中学到的整个决策树（不修剪）。

(4) 在这个未修剪树中训练集错误是什么?

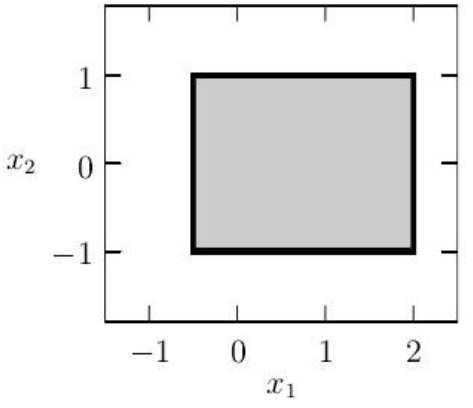
*下面的数据有助于你无需借助计算器来回答这个问题:

$\log_2 0.1 = -3.32$, $\log_2 0.2 = -2.32$, $\log_2 0.3 = -1.73$, $\log_2 0.4 = -1.32$, $\log_2 0.5 = -1$.

4. 神经网络（15 分）

设计一个两层的有阈值单元的前馈神经网络，当输入（x1,x2）在图中的灰色区域时输出为 1，否则为 0。画出网络并写出所有连接的权值和所有单元的阈值。

注意：不可以反向阈值；当大于阈值的时候单元总是给一个高的输出。



5. 上机题（40 分）

在神经网络、独立成分分析、支持向量机、增强学习方法、深度学习方法中，选择一种你感兴趣的机器学习方法，结合自己的科研实际，利用该方法解决一个实际问题，编程实现并完成一篇实验报告。报告内容包括：方法的应用背景与研究现状，基本原理，算法流程，算法实现仿真结果，结果分析与比较，给出结论。（*可自主选择程序语言，可使用 MATLAB 工具箱，代码在最后以附页形式给出。）