

重庆某双一流大学机器学习期末真题

考试时间：2023 年 6 月 13 日 14:00-16:00

命题人：hgy 组题人：hzs 审题人：czy

回忆整理：Dark Knight

第一版前言

写在前面:

1. 试卷总体比较简单, 闭卷, 不能带计算器(也没有什么计算量).
2. 复习全面一点, 90 分没有什么问题.
3. 本回忆版真题于 2020 年 12 月 17 日上午 11:56 写成, 刚吃完饭, 有点撑.
4. 计院专业课的试卷似乎都不准老师发出来, 希望有学弟学妹们能将我“回忆试卷”的习惯传承下去, 为之后的学弟学妹们做一点微小的贡献 $O(\cap \cap)O$
5. 其 余 专 业 课 的 回 忆 版 试 卷 也 许 可 以 在 github.com/VayneDuan 找到, 记得 **star & follow!**

第二版前言

写在前面:

1. 试卷总体确实比较简单, 闭卷, 不能带计算器。如果需要计算 e 或者 \log 会直接给出数据, 带入即可。
2. 机器学习主打一个报菜名, 学思想, 走流程。考试时对于数学证明的要求基本没有, 只需要记住每种模型的基本思想, 算法流程, 应用场景, 优缺点即可。用 hzs 老师的话说就是 3W(What, Way, Why)+2A(Application)。
3. 第二版的出现也是为了传承计院的传统, 希望能为学弟学妹们复习提供一些便利。

简答题(10 分*5)

1. 简述监督学习和非监督学习的概念，并至少各列举一个算法。
2. 简述欠拟合和过拟合的概念，并至少各列举一个主要解决方案。
3. 试阐述预剪枝和后剪枝的优缺点。
4. 机器学习过程中训练集和测试集的划分是不得不面对的问题，试阐述留出法、K 折交叉验证法、自助法的基本原理。
5. 谈谈你对核函数的理解（从定义、基本思想、常见核函数的选择三个方面中选择至少两个方面进行分析）。

思考题(10 分*2)

1. 写出 k 均值算法的伪代码
2. 简述 bagging 和 boosting 的基本原理，并说明它们在偏差-方差分析上的主要关注点。

计算题(30 分)

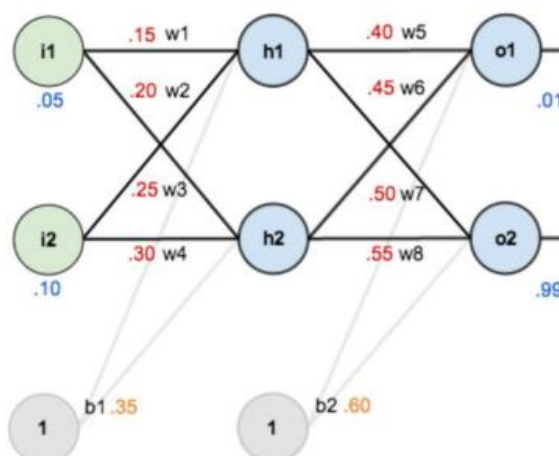
1. 给定一个 $2 \times 2 \times 2$ 的单隐层神经网络。(10 分)
 - (1) 画出这个神经网络的拓扑结构
 - (2) 进行一次前向训练，写出隐层神经元的输入和输出，以及输出层神经元的输入和输出。

输入数据 $i_1=0.05$, $i_2=0.10$;

输出数据 $o_1=0.01$, $o_2=0.99$;

初始权重 $w_1=0.15, w_2=0.20, w_3=0.25, w_4=0.30$;

$w_5=0.40, w_6=0.45, w_7=0.50, w_8=0.55$



目标：给出输入数据 i_1, i_2 (0.05和0.10)，使输出尽可能与原始输出 o_1, o_2 (0.01和0.99)接近。

hzs 老师 PPT 例题

2. 给定两位保安抓捕小偷成功率的混淆矩阵，分别计算查全率和查准率，并且说出在特定场景下哪位保安更好（希望能够抓住尽可能多的小偷）。(8分)

例：保安的能力评估

保安对进出公司的人员进行盘查，识别其中的小偷。保安有可能犯两类错误：

- 1. 把好人误认为小偷，称为假小偷（与之相反称为真小偷）。
- 2. 把小偷误认为好人，称为假好人（与之相反称为真好人）。

保安优秀与否可以按照如下的两个标准来评判：

- 1. 抓住的小偷中真小偷的比例尽量的高。
- 2. 宁可错抓好人也不放过小偷。

标准1为精准率 (Precision)

标准2为召回率 (Recall)

$$Precision = \frac{\text{抓住的真小偷数量}}{\text{抓住的真小偷数量} + \text{抓住的假小偷数量}}$$

$$Recall = \frac{\text{抓住的真小偷数量}}{\text{抓住的真小偷数量} + \text{放走的假好人数量}}$$

- 假设某公司每天进出100人，其中80个好人，20个小偷，某保安A的表现如下：

	实际上是小偷	实际上是好人	合计
抓住的小偷	5	1	6
放走的好人	15	79	94
合计	20	80	100

- 保安共计抓住6个小偷，其中5人是真小偷，另外1人是假小偷（好人被误抓），依据公式可知，该保安的 $Precision = 5 / (5 + 1) = 5 / 6 \approx 88.3\%$ 。
- 看上去这名保安抓人的准确性挺强，但他是一名好保安吗？

- 来看看召回率的计算。 $Recall = 5 / (5 + 15) = 1 / 4 = 25\%$ 。放走的假好人就是被保安误认为好人的小偷。所以 $Recall$ 的分母总是等于小偷的总数（而 $Precision$ 的分母是抓住的人的总数）。从召回率可以看出，这名保安放走了3/4的小偷，这实在不能算优秀的保安。

- 某保安B的表现如下：

	实际上是小偷	实际上是好人	合计
抓住的小偷	20	20	40
放走的好人	0	60	60
合计	20	80	100

- 根据上述公式计算， $Precision = 50\%$ ， $Recall = 100\%$ 。虽然精度不高，误抓了不少好人，但这名保安抓住了全部小偷。

hzs 老师 PPT 例题

3. 朴素贝叶斯网络模拟计算，给出样本，计算出预测结果。(12 分)

□ 某个医院早上来了六个门诊的病人，他们的情况如下表所示：

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

□ 现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

□ 根据贝叶斯定理可得：

$$P(\text{感冒}|\text{打喷嚏} \times \text{建筑工人}) = \frac{P(\text{打喷嚏} \times \text{建筑工人}|\text{感冒}) \times P(\text{感冒})}{P(\text{打喷嚏} \times \text{建筑工人})}$$

□ 根据朴素贝叶斯条件独立性的假设可知，“打喷嚏”和“建筑工人”这两个特征是独立的，因此，上面的等式就变成了

$$P(\text{感冒}|\text{打喷嚏} \times \text{建筑工人}) = \frac{P(\text{打喷嚏}|\text{感冒}) \times P(\text{建筑工人}|\text{感冒}) \times P(\text{感冒})}{P(\text{打喷嚏}) \times P(\text{建筑工人})}$$

hzs 老师 PPT 例题