**本科《机器学习》课复习重点**

**计算题：**

设在某个局部地区细胞识别中正常  和异常  两类的先验概率分别为：  
正常状态：  
异常状态：  
现有一待识别的细胞，其观察值为 ，从类条件概率密度分布曲线上查得

  
试使用贝叶斯决策对该细胞  进行分类（要求给出具体计算过程及计算结果）

**简答题：**

--试简述你对机器学习中独立同分布概念的理解。

全体样本服从一个未知分布，每个样本都是独立地从这个分布上采样获得

独立：每次抽样之间没有关系，不会相互影响

同分布：训练集和全体分布一致、相同

--试简述你对信息增益概念的理解。

在决策树算法中，信息增益是特征选择的一个重要指标，信息增益代表了在一个条件下，信息复杂度（不确定性）减少的程度

--试简述你对SVM（支持向量机）中Soft Margin（软间隔）的理解。

允许训练的模型中，部分样本（离群点或者噪音点）不必满足该约束，同时在最大化间隔时，不满足约束的样本应该尽可能的少。

--试简述什么是聚类并概要说明其中一种具体的聚类方法。

--试简述你对集成学习的理解。

**论述题：**

请论述你对深度学习的理解（提示：尽量全面和深入）

深度学习是用于建立、模拟人脑进行分析学习的神经网络，并模仿人脑的机制来解释数据的一种机器学习技术

本质上多隐层人工神经网络

属于深层模型，对于具体地任务，利用给定的一批数据，先训练一个多隐层人工神经网络，然后使用它，这就是深度学习

自动学习特征、超强的非线性建模能力

难以人工定义特征、大量标记样本、高性能计算资源

鲁棒性不足：假阳性；不可解释性；超参选择

**关于期末考试的几点说明：**

1. 所有试题，包括判断题，都请在答卷纸上写清楚题号和答案，不要直接回答在试题页上；
2. 希望大家考试都能顺利通过，因此，给出了上述复习重点，请大家提前认真准备；但是，请务必各自独立认真准备；如果发现有的同学涉及复习重点部分的答案是高度雷同的，则所有雷同卷均置为0分（因为我无法分清谁是“原创”）。因此，每位同学准备的答案，请不要给其他同学参考；此外，预先准备的关于复习重点部分的回答，可预先熟记，但不允许带入考场，咱们采取的是闭卷考试形式；
3. 第一道大题是10分的判断题；请特别注意，这10个题目是带有反扣分规则的。答对了，每题得1分；不回答，得0分；答错了，反扣1分。如果第一道大题得的总分是负分，将从试卷其他部分的得分中扣除相应分数。这是为了防止有人“蒙”。因此，建议有把握的，回答；无把握的，不答；
4. 因为该课程分数，涉及你们的研究生推免资格、奖学金的评定等，在尽可能让大家都及格的前提下，根据掌握程度的不同，分数也要有合理的区分。分数竞争空间在这样几个方面：（1）第一道大题10分的判断题；（2）部分简答题，不在上述复习重点中；（3）在复习重点中的题目，各位同学准备、回答的情况，还是有明显区别的，也能反应出大家平时学的怎么样；（4）实验分数；实验分数占总分的比例，我后边会根据实验分数的具体情况确定，尽量按有利于大家分数高一些的方向去定比例；
5. 总体来说，学习认真努力的学生，分数会高；不够努力的学生，分数会低。但，总有个例，如同高考，学习最努力的学生，未必就是分数最高的学生，也有个别学生可能学习不是那么努力，但掌握的却很好，分数高。这是学习上的天分问题。我只能根据答卷情况评定分数。