

2018.5.1

中国科学技术大学 2018-2019 第一学期  
 《机器学习与知识发现》期末考试试题

姓名: 吴为光 学号: GAI00117 成绩: \_\_\_\_\_

一、请评价两个分类器 M1 和 M2 的性能。所选择的测试集包含 26 个二值属性，记作 A 到 Z。表中是模型应用到测试集时得到的后验概率(图中只显示正类的后验概率)，因为这是二类问题，所以  $P(C) = 1 - P(\bar{C})$ ,  $P(C|A, \dots, Z) = 1 - P(\bar{C}|A, \dots, Z)$ 。  
 假设需要从正类中检测实例。(10 分)  
 1) 画出 M1 和 M2 的 ROC 曲线(画在一幅图中)。哪个模型更好？给出理由。  
 2) 对模型 M1，假设截止阈值  $t=0.5$ ，换句说话，任何后验概率大于  $t$  的测试实例都被看作正例。计算模型在此阈值下的 precision, recall 和 F-score。

实例	真实类	$P(+ A, \dots, Z, M1)$	$P(- A, \dots, Z, M2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

二、复杂的模型在训练过程中，通常会产生过拟合的现象。试从以下三种模型：  
 (1) 支持向量机，(2) 决策树，(3) 神经网络中以一种模型为例，简单说明如何避免过拟合现象。(5 分)

三、如下表数据前四列是天气情况(阴晴 outlook, 气温 temperature, 湿度 humidity, 风 windy)；最后一列是类标签，表示根据天气情况是否出去玩。(10 分)

outlook	Temperature	humidity	windy	play
sunny	Hot	high	FALSE	no
sunny	Hot	high	TRUE	no
overcast	Hot	high	FALSE	yes
rainy	Mild	high	FALSE	yes
rainy	Cool	normal	FALSE	yes
rainy	Cool	normal	TRUE	no
overcast	Cool	normal	TRUE	yes
sunny	Mild	high	FALSE	no
sunny	Cool	normal	FALSE	yes
rainy	Mild	normal	FALSE	yes

$\text{Ent}(D) = \frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}$

1) 根据上述训练集，选出属性?  
 2) 请画出两层决策树，  
 3) 使用朴素贝叶斯分类器，  
 $\text{humidity}=\text{normal}, \text{windy}=\text{false}$

四、深度神经网络的优缺点是什么？试述其优缺点。(5 分)

五、已知正例数  $n_1=10$ ，负例数  $n_2=10$ ，求误判率。  
 $(3,2)^T$ ，试求其误判率。

六、请简述核方法的分类器的实现原理。

七、K-means  
 a) 首先  
 b) 每个  
 c) 计算  
 d) 赋值  
 e) 重复

八、



1) 根据上述训练数据，基于信息增益决策树应该选哪个属性作为第一个分类属性？

2) 请画出两层决策树模型。

3) 使用朴素贝叶斯方法预测测试样本 (`outlook=rainy, temperature=cool, humidity=normal, windy=False`) 的类标号。

四、深度网络的训练突破，有一个很重要的原因是采用了新型激活函数 ReLU(如下)，试述其优缺点。(5 分)

$$y = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

五、已知正例点  $x_1 = (1, 2)^T, x_2 = (2, 3)^T, x_3 = (3, 3)^T$ ，负例点  $x_4 = (2, 1)^T, x_5 = (3, 2)^T$ ，试求最大间隔分离超平面，并指出所有的支持向量。(10 分)

六、请简述构建组合(集成)分类器的几种方法，并说明集成分类器能够改善分类器性能的原因。集成学习中多样性增强的方法有哪些？(10 分)

七、K-medoids 算法是一种聚类算法，其具体的算法流程如下：(10 分)

- 首先随机选取一组聚类样本作为中心点集
- 每个中心点对应一个簇
- 计算各样本点到各个中心点的距离(如欧几里得距离)，将样本点放入距离中心点最短的那个簇中
- 计算各簇中，距簇内各样本点距离的绝对误差最小的点，作为新的中心点
- 如果新的中心点集与原中心点集相同，算法终止；如果新的中心点集与原中心点集不完全相同，返回 b)

- 试阐述 K-medoids 算法和 K-means 算法共同的缺陷；
- 试阐述 K-medoids 算法相比于 K-means 算法的优势；
- 试阐述 K-medoids 算法相比于 K-means 算法的不足。

八、有下列五个样本，请使用 PCA 变换把特征空间维数降到一维。(10 分)

	Feature1	Feature2
Sample1	1	-1
Sample2	1	1
Sample3	2	1
Sample4	2	2
Sample5	4	2

九、试从先验概率分布的角度说明 L1 正则化和 L2 正则化各相当于假设参数服从什么分布，并解释 L1 相比于 L2 为什么容易获得稀疏解。(10 分)

十、如果概率密度函数满足：

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_1, x_3)p(x_5|x_3, x_4)$$

画出对应的贝叶斯网络，马尔可夫随机场和因子图。(10 分)



十一、现有一个城市的数据集，包括交通卡、交通事故、出租车轨迹、公交车运行、地铁运行、空气质量、气象检测、新浪微博等（具体特征如下表）。(10分)

序号	数据集名称	具体数据项
1	城市道路交通指数	状态、区域、当前指数、参考指数、指数差值
2	地铁运行数据	线路、车站、换乘站数据、首末班车各站时刻表数据、站间运行时间数据、限流车站、封站数据、路网票价矩阵、列车实时到发站台时刻、线路拥挤及阻塞数据、出入口、厕所、残疾人电梯数据
3	一卡通乘客刷卡数据	卡号、交易日期、交易时间、线路/地铁站点名称、行业名称（公交、地铁、出租、轮渡、P+R停车场）、交易金额、交易性质（非优惠、优惠、无）
4	浦东公交车实时数据	设备号码,线路编码,站点编码,协议编号,进出站状态,方向,车载上报时间、编码对应表
5	强生出租汽车行车数据	车辆ID、GPS时间、经纬度、速度、卫星颗数、营运状态高架状态、制动状态
6	空气质量状况	序号,日期,PM2.5,PM10,O3,SO2,NO2,CO,AQI,质量评价,首要污染物
7	气象数据	日期、时间、监测点、天气类型、温度、风速、风向、降水量
8	道路事故数据	事故ID、事故类型、事故地点、事故时间
9	高架匝道关闭数据	匝道ID、位置信息、关闭时间、开放时间
10	新浪微博交通数据	涵盖路况、交通工具、天气等与交通相关的关键词的微博信息

请利用你所学过的机器学习和数据挖掘的方法解决预测该城市空气质量的问题：

- (1) 哪些数据或者特征可能用到，并简要说明原因
- (2) 可以使用所学过的哪些机器学习方法解决该问题？
- (3) 请简要给出一个解决方案（最大限度地利用现有数据）。