2022-2023下机器学习基础期末试题

2023年6月12日16：10-18:10

L.C.

一、名词解释（20分）

1. **无监督学习**

   无监督学习：数据 X 已知，标签 y 未知。常见的无监督学习应用有聚类算法、异常检测等。

2. **代价敏感学习**

   代价敏感学习是一种机器学习方法，它考虑到不同的分类错误可能会带来不同的代价。在代价敏感学习中，我们为每种分类错误都分配一个代价，并且在训练模型时优化这些代价的总和，而不是简单地优化分类准确率。

3. **核函数**

   初衷：当原始空间下线性不可分时候，我们希望将数据映射到高维空间使之线性可分。我们使用$\Phi(x_i)$为函数，将低维的样本x_i映射到高维。但是我们在求解目标函数对应的参数时候，必须用到两个样本的内积，而**在高维空间中计算两个样本的内积计算难度较大。**

   因此我们需要有一个函数，可以避免在高维空间计算$\Phi(x)^T\Phi(x)$，但是又可以得到内积的结果，这就是核函数。

4. **马尔科夫性**

   当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么此随机过程即具有**马尔可夫性质。**

二、简答题（60分）

1. **简述留出法和交叉验证法。（10分）**

   ① **留出法：**将数据集D划分为两个互斥的集合，其中一个饿集合作为训练集S，另一个作为测试集T，在S上训练出模型后，用T来评估其测试误差，作为对泛化误差的估计。需要注意保持数据分布一致性、多次重复划分取平均、测试集不能太大、不能太小

   ② **交叉验证法：**将数据集D划分为k个大小相似的互斥子集，每个子集都尽可能保持数据分布的一致性，每次用k-1个子集的并集作为训练集，余下的那个子集作为测试集。

2. **简述KNN的基本思想，如何初始化K。（10分）**

   ① KNN被认为是一种懒惰的学习算法，它根据数据集与邻居的相似性来进行分类

   ② “K”代表分类时考虑的数据集项目的数量

   ③ K-最近的邻居被称为非参数化方法，与其他监督学习算法不同，K-最近的邻居并不从训练数据中学习明确的映射f。它只是在测试时使用训练数据来进行预测（利用训练数据选取K个点来分类）

   ① 应当选择奇数，保证多数表决时100%能产生结果

   ② 1-NN表现通常都不错

   ③ K往往小于训练样本总数的1/2次方

   ④ 可以使用交叉验证来评判K值

3. **简述偏差和方差的概念。（10分）**

   偏差：学习算法中错误假设造成的误差。偏差量度了学习算法的预期期望与真实结果的偏离程度，即刻画了学习算法本身的拟合能力。方差：灵敏度对训练集中小波动的误差。方差量度了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响。

4. **列出SVM的目标函数。（10分）**

   找到一个超平面使得尽可能将两类数据分离开，使得到超平面最近的点的距离（也就是间隔）最大。

$$\max_{w,b} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1,2,\ldots m$$

$$\downarrow$$

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1,2,\ldots m$$

5. **简述线形回归的基本思想，如何求解线性回归方程。（10分）**

   “线性回归”试图学得一个**线性拟合函数**以尽可能地拟合数据，并尽可能准确地预测数据。

general model: $\boxed{f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \ldots + w_dx_d + b}$

$x_1, x_2, \ldots, x_d$: the feature of sample

$w_1, w_2, \ldots, w_d$: weight, represent the importance of corresponding feature

w 直观表达了各属性在预测中的重要性，因而线性模型具有很好的可解释性。

general model: $f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \ldots + w_dx_d + b$

vector formal: $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$

$$\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$$

$$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$$

6. **简述线性判别法（LDA），分析优劣势。（10分）**

   LDA是Fisher线性判别式的推广，Fisher线性判别法是一种用于统计学、模式识别和机器学习的方法，用于寻找表征或分离两类或两类以上对象或事件的特征的线性组合。

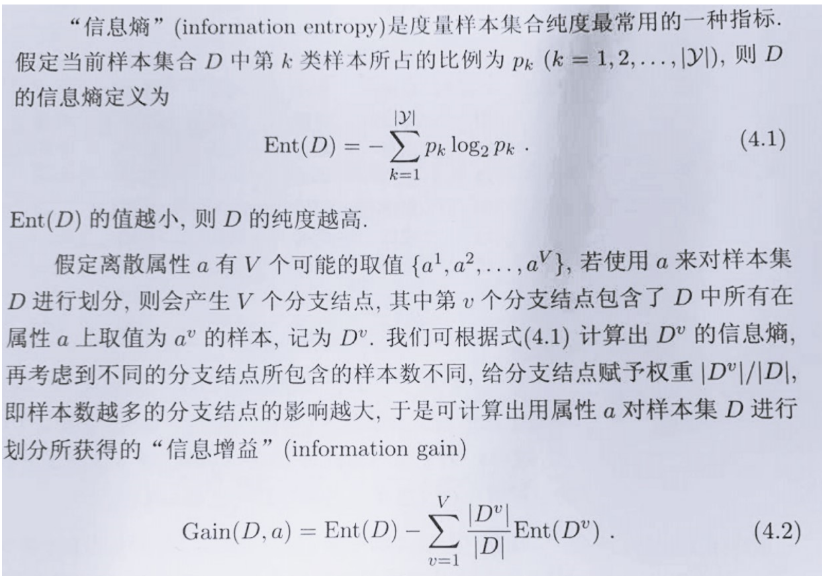
   优点：监督学习；缺点：对于具有C类的数据，只得到具有（C-1）维LDA的特征，要求数据符合高斯分布，容易导致过拟合。

三、综合分析题（20分）

	1	2	3	4	5	6	7	8	9	0
属性1	A	A	C	A	B	C	C	B	A	C
属性2	▲	▲	■	●	▲	●	■	■	■	●
	0	0	0	0	1	1	1	1	1	1

N	0.25	0.33	0.4	0.6	0.67	0.75
$log_2 N$	-2	-1.6	-1.32	-0.73	-0.57	-0.41

1. **给定10个样本和属性，如上表所示。从信息增益的角度考虑，判断决策树应选择哪个属性，给出具体计算过程。**



“信息熵”(information entropy)是度量样本集合纯度最常用的一种指标. 假定当前样本集合 D 中第 k 类样本所占的比例为 p_k $(k=1,2,\ldots,|\mathcal{Y}|)$, 则 D 的信息熵定义为

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \ . \tag{4.1}$$

$\mathrm{Ent}(D)$ 的值越小, 则 D 的纯度越高.

假定离散属性 a 有 V 个可能的取值 $\{a^1, a^2, \ldots, a^V\}$, 若使用 a 来对样本集 D 进行划分, 则会产生 V 个分支结点, 其中第 v 个分支结点包含了 D 中所有在属性 a 上取值为 a^v 的样本, 记为 D^v. 我们可根据式(4.1) 计算出 D^v 的信息熵, 再考虑到不同的分支结点所包含的样本数不同, 给分支结点赋予权重 $|D^v|/|D|$, 即样本数越多的分支结点的影响越大, 于是可计算出用属性 a 对样本集 D 进行划分所获得的“信息增益”(information gain)

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \ . \tag{4.2}$$

2. **两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大?**

   分清和定义现象和规律。这儿有两个规律：从一号碗来规律，从二号碗来规律。
   有两种现象：水果糖现象，巧克力现象。而且知道两个一模一样的碗，所以两个规律的概率一样，P(从一号碗来规律) = P(从二号碗来规律) = 0.5。同时知道P(水果糖现象|从一号碗来规律) = 30/(30+10)=0.75，
   P(巧克力现象|从一号碗来规律)=10/(30+10)=0.25；P(水果糖现象|从二号碗来规律)=20/(20+20)=0.5，
   P(巧克力现象|从二号碗来规律)=20/(20+20)=0.5。另外，P(水果糖现象) = (30+20)/(30+10+20+20)=0.625，
   P(巧克力现象)=(10+20)/(30+10+20+20)=0.375。
   现在的问题是观察到了一个水果糖现象，要求推断后面的规律，
   即从一号碗来的规律的概率是多大，
   也就是P(从一号碗来规律|水果糖现象)。

   P(从一号碗来规律|水果糖现象) = P(水果糖现象|从一号碗来规律)*P(从一号碗来规律)/P(水果糖现象)=0.75*0.5/0.625=0.6。

评价：老师wqc，上课压根不用听，考试是第一年闭卷，需要背的内容相当多。考试出乎意料，和往年题相似度较低，但是计算都是原题，考了一些边角的内容，平时分40的情况下有点困难。