北京大学数学科学学院期末试题

2019-2020学年第2学期

考试科目：机器学习基础　　考试时间：2020年6月19日8:30-10:30

学号：　　姓名:

说明：**共七道大题, 请将答案写在纸上, 拍照上传。**

凡计算、推导等都需要给出具体过程；切勿向外泄露或者发布考题。

1. (20分)在下述隐马尔可夫模型λ中，可能的观测值集合为$\{\nu_1, \nu_2, \nu_3\}$, 可能的状态集为$\{q_1, q_2, q_3\}$, 状态转移概率矩阵为$A = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.5 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$, 观测概率矩阵为$B = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$, 初始状态概率向量为$\pi = (0.6, 0.2, 0.2)^T$, 观察序列$O = (\nu_3, \nu_2, \nu_1)$。

   (1) 用**后向算法**计算$P(O|\lambda)$.

   (2) 用**维特比算法**求最优状态序列.

   (3) 以上述模型λ作为$\lambda^{(n)}$,以**Baum-Welch** 算法计算$\pi_2^{(n+1)}$和$a_{13}^{(n+1)}$.

2. (20分)考虑如表 1所示的二类分类问题数据，类别$y \in \{\text{Yes}, \text{No}\}$。

表 1: 数据集D

	英语(E)	成绩(S)	科研经历(R)	CLASS:奖学金(FS)
1	A	EXCELLENT	Y	Yes
2	B	EXCELLENT	N	Yes
3	A	FAIR	Y	No
4	B	FAIR	N	No
5	A	GOOD	N	No
6	B	GOOD	Y	Yes
7	B	EXCELLENT	Y	Yes
8	A	FAIR	N	No
9	B	GOOD	N	No
10	C	EXCELLENT	N	No
11	A	GOOD	Y	Yes
12	C	EXCELLENT	Y	Yes
13	C	FAIR	Y	No
14	C	GOOD	N	No
15	C	GOOD	Y	Yes

(1) 利用朴素贝叶斯法对数据

$$x_p = (\mathrm{B}, \mathrm{EXCELLENT}, \mathrm{N})^T, \quad x_q = (\mathrm{B}, \mathrm{FAIR}, \mathrm{N})^T$$

的类别分别进行预测。

(2) 利用**CART**算法生成决策树（不考虑剪枝）。

3. (25分) 设训练数据集为$\{(x_i, y_i)\}_{i=1}^N$，其中$x_i \in \mathcal{X} \subseteq \mathbf{R}^n$，$y_i \in \mathcal{Y} = \{+1, -1\}$, $i = 1, 2, \cdots, N$. 基于此训练数据集，

(1) 分别给出**线性支持向量机**的原始最优化问题、对偶问题和以合页损失函数表示的形式.

(2) 写出**高斯径向基函数**分类器(支持向量机)所对应的最优化问题和分类决策函数。

(3) 考虑以拉格朗日乘子给出的如下最优化问题：

$$\begin{aligned} &\min_{\alpha,b,\xi} \frac{1}{2} \parallel \alpha \parallel^2 + C(\sum_{i=1}^N \xi_i) \\ s.t. \ \ & y_i \left(\sum_{j=1}^N \alpha_j y_j x_j \cdot x_i + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & \alpha_i \geq 0, \ i = 1, 2, \cdots, N. \end{aligned}$$

(a) 论证除了关于拉格朗日乘子的非负约束之外，在形式上此优化问题与某个支持向量机的原始最优化问题一致。

(b) 推导上述关于拉格朗日乘子的最优化问题的对偶问题。

4. (10分) 若学习算法$\mathcal{L}$基于训练样本集$D = \{(x_i, y_i)\}_{i=1}^N$学到的假设为$h_D$.

(1) 给出损失函数为0-1损失时的留一(Leave-One-Out)损失。

(2) 若D是线性可分的, 讨论学到的支持向量机的支持向量个数与留一损失之间的关系.

(3) 如果$x_i \in \mathbf{R}^n$，结合计算学习理论知识（如Rademacher复杂度，VC维等），给出基于$D = \{(x_i, y_i)\}_{i=1}^N$学到的线性支持向量机模型的一个泛化误差界，并简要分析一下所给出的界与N和n的关系。

5. (15分) 对岭回归、Lasso算法和支持向量回归模型的优缺点进行分析比较, 并基于分析为其中某个模型设计一个**新**变体，简要论证其合理性。

6. (5分)给出你自己对**PAC可学习**和**不可知PAC可学习**的理解。

7. (5分)列出两条你认为机器学习中比较重要（或比较有用，或比较有启发性的）基本思想（或策略或技巧）, 并简要阐述理由。