机器学习期末考试

2021 春季学期

姓名：　　学号：

注意事项：

1. 本试卷共**六大题**，总分**100分**，考试时间为**120分钟**。
2. 请在每一题题后进行作答，最后三页为草稿纸。
3. 使用黑色签字笔作答，确保字迹清晰。
4. 本次考试除计算器外，禁止随身携带任何电子设备。

一. 学习理论（8+5+5=18分）

(1) 根据Hoeffding不等式（ϵ为给定误差，N为样本容量）：

$$\mathbb{P}[\,|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N},$$

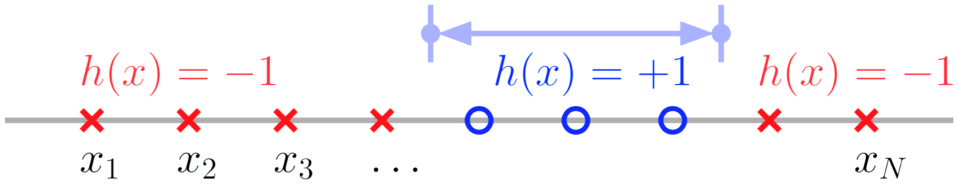
推导出$\mathscr{H}$只有有限个元素($|\mathscr{H}| < +\infty$)的时候的泛化误差(generalization error)。

(2) 根据VC不等式

$$\mathbb{P}[\sup_{h\in\mathscr{H}} |E_{in}(h) - E_{out}(h)| > \epsilon] \leq 4m_{\mathscr{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N},$$

讨论影响模型泛化性能的因素。其中$m_{\mathscr{H}}(N)$为$\mathscr{H}$的增长函数(growth function)。

(3) 考虑一维情况下线段分类器，即$h(x) = \{x \in [a,b]\}$。给出其增长函数$m_{\mathscr{H}}(N)$和VC维度d_{vc}。



二. 极大似然估计和极大后验估计（MLE and MAP）（8+5=13分）：

假设所估计参数 θ 满足密度分布$p(\theta)$，根据Bayes定理可得后验(posterior)为

$$p(\theta \mid y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta).$$

现在考虑待估计的参数为$\theta \in \mathbb{R}^n$，假设数据集为$\{(x_i, y_i)\}_{i=1}^m$， θ*是极大化后验概率的最优解，那么

$$\begin{aligned}\theta^* &= \arg\max_\theta \prod_{i=1}^m p(y_i|x_i,\theta)p(\theta) \\ &= \arg\max_\theta \sum_{i=1}^m \log(p(y_i|x_i,\theta)) + \log p(\theta)\,.\end{aligned}$$

(1) 现假定$y_i = x_i^T\beta + \epsilon_i$，其中 $x_i, \beta \in \mathbb{R}^n, \epsilon \sim \mathcal{N}(0,\sigma^2)$，$\beta_j \sim \mathcal{N}(0,\frac{1}{2\lambda})$，假定 x_i为给定数值，σ, λ为已知。写出估计 β 的MAP的目标函数。提示: 如果 $z \sim \mathcal{N}(\mu, \sigma^2)$, 则其分布函数

$$p(z) \propto \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

(2) 写出问题 (1) 最优解的解析表达式。

三. 优化基础（5+5+8=18分）

(1) 对于优化问题 $\min_{x\in\mathbb{R}^n} f(x)$，其中$f(x)$是强凸(strongly convex)的。证明牛顿方向是下降方向（descent direction）。

(2) 给出拟牛顿法（quasi-Newton）中 Hessian近似矩阵所必须满足的割线方程（Secant Equation）。

(3) 现在采取单位矩阵的倍数 αI 来近似Hessian矩阵，得到超定方程。给出此时最小二乘意义下Secant Equation的最优解 α。

四. 支撑向量机（8+5=13分）

假设在数据集$\{(x_i, y_i)\}_{i=1}^n$，$x_i \in \mathbb{R}^n$, $y_i \in \{+1, -1\}$ 上训练硬边界支撑向量机（hard-margin SVM）得到最优的模型为$f(x; w_*, b_*)$。

(1) 写出如下L1-SVM的Lagrange对偶问题：

$$\begin{aligned} \min_{w} \quad & \|w\|_1 \\ \text{s.t} \quad & y_i(w^T x_i + b) \geq 1,\ i = 1,...,n\ . \end{aligned}$$

(2) 现考虑0-1损失函数 :

$$L(y_i, f(x_i; w_*, b_*)) = \begin{cases} 0, y_i = f(x_i; w_*, b_*) \\ 1, y_i \neq f(x_i; w_*, b_*) \end{cases}$$

证明所得到的分类器$f(x; w_*, b_*)$ 的"留一交互验证误差" E_{cv}（leave-one-out cross validation error）满足：$E_{cv} \leq \frac{n_{sv}}{n}$，其中 n_{sv} 为支撑向量的个数。

五. 决策树（8分）

考虑如下数据集，其特征和类别如下表所示：

数据集		特征	
数据	类别	A1	A2
1	+	T	T
2	+	T	T
3	–	T	F
4	+	F	F
5	–	F	T
6	–	F	T

给出ID3算法构建决策树根节点的计算过程。（提示: $\log_2 3 = 1.5850$）

六. 阐述题（每题5分，共30分）：
用自己的语言阐述以下问题。
(1) 什么是期望风险极小化、经验风险极小化、以及结构风险极小化。

(2) 阐述在梯度下降法里面，关于学习率，你如何选（注意区分小规模，大规模问题）。

(3) 在利用正则化来克服过拟合时，如何选取合适的正则化参数?

(4) 现考虑在训练时通过正则化问题

$$\min_{w} \quad \text{loss}(w) + \lambda R(w)$$

来选取最佳模型，其中$R(w)$为正则项。假如计算时间足够长，第一种策略是在

$$\lambda \in \{0,\ 10^{-6},\ 10^{-5}, \ldots,\ 10^{-1},\ 1,\ 10\}$$

得待选取值里选取表现最佳的λ；而第二种是在

$$\lambda \in \{0, 10^{-10},\ 2\times 10^{-10},\ 3\times 10^{-10}, \ldots, 10\}$$

里选取最佳的λ。你认为哪种策略选取得到的λ所对应的模型性能更佳？为什么?

(5) 阐述logistic regression、SVM、decision tree、DNN这四类分类器的优缺点。

(6) 对于方差-偏差分解公式：

$$\mathbb{E}_{\mathcal{D},\epsilon}[(y^* - h(x^*))^2] = \text{variance} + \text{bias}^2 + \text{noise}\,.$$

其中，$y = f(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned} \text{variance} &= \mathbb{E}_{\mathcal{D}}[(h(x^*) - \bar{h}(x^*))^2] \\ \text{bias}^2 &= [\bar{h}(x^*) - f(x^*)]^2 \\ \text{noise} &= \sigma^2 \end{aligned}$$

分析其各因素对泛化性能的影响。