

机器学习导论 习题三

学号, 姓名, 邮箱

2023 年 4 月 25 日

作业提交注意事项

1. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱;
2. 本次作业需提交作答后的该 pdf 文件、编程题代码 (.py 文件); **请将二者打包为 .zip 文件上传**. 注意命名规则, 三个文件均命名为“学号_姓名”+ “.后缀” (例如 “211300001_张三” + “.pdf”、“.py”、“.zip”);
3. 若多次提交作业, 则在命名 .zip 文件时加上版本号, 例如 “211300001_ 张三_v1.zip” (批改时以版本号最高的文件为准);
4. 本次作业提交截止时间为 **5 月 2 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, **作业命名不规范**, 将会被扣除部分作业分数; 除特殊情况 (如因病缓交, 需出示医院假条) 逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
5. 本次作业提交地址为 [here](#), 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.

1 [20pts] Representer Theorem

表示定理告诉我们, 对于一般的损失函数和正则化项, 优化问题的最优解都可以表示为核函数的线性组合. 我们将尝试证明表示定理的简化版本, 并在一个实际例子中对其进行应用. 请仔细阅读《机器学习》第六章 6.6 节, 并回答如下问题.

- (1) [10pts] 考虑通过引入核函数来将线性学习器拓展为非线性学习器, 优化目标由结构风险和经验风险组成:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\mathbf{w}^T \phi(\mathbf{x}_i), y_i) + \frac{\lambda}{2} \|\mathbf{w}\|^2,$$

其中映射 $\phi: \mathcal{X} \rightarrow \mathbb{H}$ 将样本映射到特征空间 \mathbb{H} , \mathcal{L} 为常见的损失函数, 并记 $\mathbf{X} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_m)]$ 为映射后的数据矩阵. 请证明: 优化问题的最优解 \mathbf{w}^* 属于矩阵 \mathbf{X} 的列空间, 即 $\mathbf{w}^* \in \mathcal{C}(\mathbf{X})$.

(提示: 给定线性子空间 \mathcal{S} , 任意向量 \mathbf{u} 有唯一的正交分解 $\mathbf{u} = \mathbf{v} + \mathbf{s} (\mathbf{v} \in \mathcal{S}, \mathbf{s} \in \mathcal{S}^\perp)$. 你需要选取合适的线性子空间, 对 \mathbf{w} 进行正交分解)

- (2) [10pts] 在核岭回归问题 (KRR, kernel ridge regression) 中, 优化目标为:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^m (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2.$$

根据第一问的结论, 该优化问题的最优解满足 $\mathbf{w}_{\text{KRR}}^* = \mathbf{X}\boldsymbol{\alpha}$. 请给出此处 $\boldsymbol{\alpha}$ 的具体形式. 值得一提的是, $\boldsymbol{\alpha}$ 是 KRR 问题对偶问题的最优解.

(提示: 你需要先求出 $\mathbf{w}_{\text{KRR}}^*$ 的具体形式)

Solution. 此处用于写解答 (中英文均可)

2 [20pts] Leave-One-Out error in SVM

《机器学习》第 2.2.2 节中我们接触到了留一法 (Leave-One-Out), 使用留一损失作为分类器泛化错误率的估计, 即: 每次将一个样本作为测试集, 其余样本作为训练集, 最后对所有的测试误差取平均. 对于 SVM 算法 \mathcal{A} , 令 h_S 为该算法在训练集 S 上的输出, 则 \mathcal{A} 的经验留一损失可形式化为

$$\hat{R}_{\text{LOO}}(\mathcal{A}) = \frac{1}{m} \sum_{i=1}^m 1_{h_{S \setminus \{\mathbf{x}_i\}}(\mathbf{x}_i) \neq y_i}.$$

本题将通过探索留一损失的一些数学性质, 分析 SVM 泛化误差与支持向量个数的联系, 并给出一个期望意义下的泛化误差界. (注: 本题仅考虑可分情形, 即数据集是线性可分的)

- (1) [5pts] 在实际应用中, 测试误差相比于泛化误差是很容易获取的. 我们往往希望测试误差是泛化误差较为准确的估计, 至少应该是无偏估计. 试证明留一损失是数据集大小为 $m - 1$ 时泛化误差的无偏估计, 即

$$\mathbb{E}_{S \sim \mathcal{D}^m} [\hat{R}_{\text{LOO}}(\mathcal{A})] = \mathbb{E}_{S' \sim \mathcal{D}^{m-1}} [R(h_{S'})].$$

- (2) [5pts] SVM 的最终模型仅与支持向量有关, 支持向量完全刻画了决策边界. 这一现象可以抽象表示为, 如果样本 \mathbf{x} 并非 h_S 的支持向量, 则移除该样本不会改变 SVM 模型, 即 $h_{S \setminus \{\mathbf{x}\}} = h_S$. 这一性质在分析误差时有关键作用, 考虑如下问题: 如果 \mathbf{x} 不是 h_S 的支持向量, $h_{S \setminus \{\mathbf{x}\}}$ 会将 \mathbf{x} 正确分类吗, 为什么? 该问题的结论的逆否命题是什么?

- (3) [10pts] 基于上一小问的结果, 试证明下述 SVM 的泛化误差界限:

$$\mathbb{E}_{S \sim \mathcal{D}^m} [R(h_S)] \leq \mathbb{E}_{S \sim \mathcal{D}^{m+1}} \left[\frac{N_{SV}(S)}{m+1} \right],$$

其中 $N_{SV}(S)$ 为模型 h_S 支持向量的个数. 从这一泛化误差界中, 我们能够看到 SVM 的泛化能力与支持向量个数之间有紧密的联系.

Solution. 此处用于写解答 (中英文均可)

3 [30pts] Margin Distribution

SVM 的核心思想是最大化最小间隔, 以获得最鲁棒的分类决策边界. 然而, 近年来的一些理论研究表明, 最大化最小间隔并不一定会带来更好的泛化能力, 反而优化样本间隔的分布可以更好地提高泛化性能. 为了刻画间隔的分布, 我们可以使用样本间隔的一阶信息和二阶信息, 即间隔均值和间隔方差.

给定训练数据集 $\mathcal{S} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, $\phi: \mathcal{X} \rightarrow \mathbb{H}$ 为映射函数, 我们记 $\mathbf{X} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_m)]$ 为映射后的数据矩阵, $\mathbf{y}^T = [y_1, \dots, y_m]$ 为标签向量, \mathbf{Y} 是对角元素为 y_1, \dots, y_m 的对角矩阵. 请回答如下问题.

- (1) [5pts] 间隔均值与间隔方差分别定义为:

$$\gamma_m = \frac{1}{m} \sum_{i=1}^m y_i \mathbf{w}^T \phi(\mathbf{x}_i),$$

$$\gamma_v = \frac{1}{m} \sum_{i=1}^m (y_i \mathbf{w}^T \phi(\mathbf{x}_i) - \gamma_m)^2.$$

请使用题给记号, 化简上述表达式.

- (2) [5pts] 考虑标准的软间隔 SVM(课本公式 (6.35)) 且引入核函数. 现在, 我们希望在基础上进行改进: 最大化样本间隔的均值, 并且最小化样本间隔的方差. 令间隔均值的相对权重为 μ_1 , 间隔方差的相对权重为 μ_2 , 请给出相应的优化问题.
- (3) [20pts] 第二问中的想法十分直接, 但是由于优化问题中的目标函数形式较为复杂, 导致对偶问题难以表示. 借鉴 SVM 中固定最小间隔为 1 的思路, 我们固定间隔均值为 $\gamma_m = 1$, 每个样本 (\mathbf{x}_i, y_i) 的间隔相较于均值的偏移为 $|y_i \mathbf{w}^T \phi(\mathbf{x}_i) - 1|$. 此时仅需最小化间隔方差, 相应的优化问题为

$$\min_{\mathbf{w}, \xi_i, \epsilon_i} \quad \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{m} \sum_{i=1}^m (\xi_i^2 + \epsilon_i^2)$$

$$\text{s.t.} \quad y_i \mathbf{w}^T \phi(\mathbf{x}_i) \geq 1 - \xi_i, y_i \mathbf{w}^T \phi(\mathbf{x}_i) \leq 1 + \epsilon_i, \forall i.$$

其中 $C > 0$ 为正则化系数, ξ_i 和 ϵ_i 为松弛变量, 刻画了样本相较于均值的偏移程度. 进一步地, 我们借鉴支持向量回归 (SVR) 中的做法, 引入 θ -不敏感损失函数, 容忍偏移小于 θ 的样本. 同时, 间隔均值两侧的松弛程度可有所不同, 使用参数 μ 进行平衡. 最终我们得到了最优间隔分布机 (Optimal margin Distribution Machine) 的优化问题:

$$\min_{\mathbf{w}, \xi_i, \epsilon_i} \quad \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{m} \sum_{i=1}^m \frac{\xi_i^2 + \mu \epsilon_i^2}{(1 - \theta)^2}$$

$$\text{s.t.} \quad y_i \mathbf{w}^T \phi(\mathbf{x}_i) \geq 1 - \theta - \xi_i$$

$$y_i \mathbf{w}^T \phi(\mathbf{x}_i) \leq 1 + \theta + \epsilon_i, \forall i.$$

试推导该问题的对偶问题, 要求详细的推导步骤. (提示: 借助题干中的记号, 将该优化问题表达成矩阵的形式. 你也可以引入额外的记号)

Solution. 此处用于写解答 (中英文均可)

4 [30pts] Classification Models

编程实现不同的分类算法, 并对比其表现. 详细编程题指南请参见链接: [here](#).

- (1) 请填写下表, 记录不同模型的精度与 AUC 值. (保留 4 位小数)
 - (2) 请将绘制好的, 不同模型在同一测试数据集上的 ROC 曲线图放在此处.
- 再次提醒, 请注意加入图例.

Solution. 此处用于写解答 (中英文均可)

(1) 不同模型的精度与 AUC 值记录

表 1: 不同模型的精度、AUC 值

<div>模型 指标</div>	Logistic Regression	Decision Tree	SVM
acc. on train			
acc. on test			
AUC on test			

(2) 不同模型在测试数据集上的 ROC 曲线