

# 机器学习导论 习题三

211300044, 吴羽珩, 211300044@smail.nju.edu.cn

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)

- (1) 给定线性子空间  $\mathcal{S}$ , 任意向量  $\mathbf{u}$  有唯一的正交分解  $\mathbf{u} = \mathbf{v} + \mathbf{s} (\mathbf{v} \in \mathcal{S}, \mathbf{s} \in \mathcal{S}^\perp)$  所以将  $\mathbf{w}$  在  $\mathcal{C}(\mathbf{X})$  上正交分解得到  $\mathbf{w} = \mathbf{v} + \mathbf{s}$ , 其中  $\mathbf{v} \in \mathcal{C}(\mathbf{X}), \mathbf{s} \in (\mathcal{C}(\mathbf{X}))^\perp$  由  $\mathbf{s} \in (\mathcal{C}(\mathbf{X}))^\perp$  可知, 对于  $\forall \phi(\mathbf{x}_i) \in \mathcal{C}(\mathbf{X})$ , 均有  $\mathbf{s}^\top \phi(\mathbf{x}_i) = 0$ .

所以

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\mathbf{w}^\top \phi(\mathbf{x}_i), y_i) &= \frac{1}{m} \sum_{i=1}^m \mathcal{L}((\mathbf{v}^\top + \mathbf{s}^\top) \phi(\mathbf{x}_i), y_i) \\ &= \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\mathbf{v}^\top \phi(\mathbf{x}_i) + \mathbf{s}^\top \phi(\mathbf{x}_i), y_i) \\ &= \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\mathbf{v}^\top \phi(\mathbf{x}_i), y_i) \end{aligned}$$

又因为  $\|\mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{s}\|^2 \geq \|\mathbf{v}\|^2$  (当  $\mathbf{s} = \vec{0}$  时取等), 所以我们一定有  $J(\mathbf{w}) \geq J(\mathbf{v})$  设优化目标  $\min_{\mathbf{w}} F(\mathbf{w})$  最优解为  $\mathbf{w}^*$ , 同样有  $J(\mathbf{w}^*) \geq J(\mathbf{v}^*)$  而根据最优解的定义:  $\mathbf{w}^*$  最小化了  $F(\mathbf{w})$ , 所以又有  $F(\mathbf{w}^*) \leq F(\mathbf{v}^*)$  所以得到  $F(\mathbf{w}^*) = F(\mathbf{v}^*)$ , 即  $\mathbf{s}^* = \mathbf{0}$ , 因此  $\mathbf{w}^* = \mathbf{v}^* \in \mathcal{C}(\mathbf{X})$

(2) 首先, 对原问题进行转化, 将  $F(\mathbf{w})$  转化为向量和矩阵的形式.

$$\begin{aligned} F(\mathbf{w}) &= \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^m (\mathbf{w}^T \phi(\mathbf{x}_i) - y_i)^2 \\ &= \lambda \mathbf{w}^T \mathbf{w} + (\mathbf{X}^T \mathbf{w} - \mathbf{y})^T (\mathbf{X}^T \mathbf{w} - \mathbf{y}) \end{aligned}$$

显然,  $F(\mathbf{w})$  是凸函数, 下面求解  $\mathbf{w}^*$ , 将  $F(\mathbf{w})$  对  $\mathbf{w}$  求导得:

$$2\lambda \mathbf{w} + 2\mathbf{X}\mathbf{X}^T \mathbf{w} - 2\mathbf{X}\mathbf{y}$$

令其为 0 解得:

$$\mathbf{w}^* = (\lambda \mathbf{I}_m + \mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y}$$

注意到矩阵  $\mathbf{X}^{-1}$  不一定存在, 故不能简单的左右同时左乘  $\mathbf{X}^{-1}$

进一步转化

$$\mathbf{w}^* = -(\mathbf{I}_m - (-\mathbf{X})(\lambda^{-1} \mathbf{I}_n) \mathbf{X}^T)^{-1} (-\mathbf{X})(\lambda \mathbf{I}_n)^{-1} \mathbf{y}$$

其中  $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{I}_m \in \mathbb{R}^{m \times m}, \mathbf{I}_n \in \mathbb{R}^{n \times n}$

由结论

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \mathbf{B}\mathbf{D}^{-1} = \mathbf{A}^{-1} \mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}$$

可得

$$-(\mathbf{I}_m - (-\mathbf{X})(\lambda^{-1} \mathbf{I}_n) \mathbf{X}^T)^{-1} (-\mathbf{X})(\lambda \mathbf{I}_n)^{-1} = -\mathbf{I}_m^{-1} (-\mathbf{X}) ((\lambda \mathbf{I}_n) - \mathbf{X}^T \mathbf{I}_m^{-1} (-\mathbf{X}))^{-1}$$

所以

$$\mathbf{w}^* = -\mathbf{I}_m^{-1} (-\mathbf{X}) ((\lambda \mathbf{I}_n) - \mathbf{X}^T \mathbf{I}_m^{-1} (-\mathbf{X}))^{-1} \mathbf{y} = \mathbf{X}(\lambda \mathbf{I}_n + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}$$

也即

$$\boldsymbol{\alpha} = (\lambda \mathbf{I}_n + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}$$

## 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)

- (1) 泛化误差 (Generalization Error) 即模型在总体上的误差.

注意到样本是独立的, 所以有  $\mathbb{E}(\sum X_i) = \sum \mathbb{E}(X_i)$

$$\begin{aligned} \mathbb{E}_{S \sim \mathcal{D}^m} [\hat{R}_{\text{LOO}}(\mathcal{A})] &= \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{S \sim \mathcal{D}^m} [1_{h_{S \setminus \{\mathbf{x}_i\}}(\mathbf{x}_i) \neq y_i}] \\ &= \mathbb{E}_{S \sim \mathcal{D}^m} [1_{h_{S \setminus \{\mathbf{x}_1\}}(\mathbf{x}_1) \neq y_1}] \quad (\text{因为样本是同分布的}) \\ &= \mathbb{E}_{S' \sim \mathcal{D}^{m-1}} [R(h_{S'})] \quad (S' \leftarrow S \setminus \{\mathbf{x}_1\}) \end{aligned}$$

- (2) 能够正确分类. 原因如下:

由书上最终推导出的结论 (KKT 条件):

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 \geq 0 \\ \alpha_i (y_i f(\mathbf{x}_i) - 1) = 0 \end{cases}$$

对任意训练样本  $(\mathbf{x}_i, y_i)$ , 总有  $\alpha_i = 0$  或  $y_i f(\mathbf{x}_i) = 1$ 。当  $\alpha_i = 0$  时, 样本点  $(\mathbf{x}_i, y_i)$  不会影响最终模型, 当  $y_i f(\mathbf{x}_i) = 1$  时, 所对应的样本点位于最大间隔边界上, 是一个支撑向量。所以只要原模型能正确分类, 删除任一非支撑向量, 仍然能正确分类。

逆否命题仍为真: 若  $h_{S \setminus \{\mathbf{x}\}}$  不能正确分类, 那么  $\mathbf{x}$  是  $h_S$  的支持向量。

(3) 根据 (1) 得到的结论

$$\mathbb{E}_{S \sim \mathcal{D}^m} [R(h_S)] = \mathbb{E}_{S \sim \mathcal{D}^{m+1}} [\hat{R}_{\text{LOO}}(\mathcal{A})] = \mathbb{E}_{S \sim \mathcal{D}^{m+1}} [1_{h_{S \setminus \{\mathbf{x}_i\}}(\mathbf{x}_i) \neq y_i}]$$

实际上  $\mathbb{E}_{S \sim \mathcal{D}^{m+1}} [1_{h_{S \setminus \{\mathbf{x}_i\}}(\mathbf{x}_i) \neq y_i}]$  即为  $S \sim \mathcal{D}^{m+1}$  当中某元素  $\mathbf{x}_i$  不能被  $h_{S \setminus \{\mathbf{x}_i\}}$  正确分类的概率, 根据 (2) 中得到的结论的逆否命题所以  $\mathcal{D}^{m+1}$  当中不能被正确分类的元素个数不大于支持向量的个数, 即只有从  $m+1$  个数据中选到支持向量的概率, 由此可得概率不大于  $\frac{N_{SV}(S)}{m+1}$

所以

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

### 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  $\mu_1$ , 间隔方差的相对权重为  $\mu_2$ , 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,  $\xi_i$  和  $\epsilon_i$  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  $\theta$ -不敏感损失函数, 容忍偏移小于  $\theta$  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  $\mu$  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)

(1)

$$\begin{aligned}
\gamma_m &= \frac{1}{m} \mathbf{w}^\top \mathbf{X} \mathbf{y} \\
\gamma_v &= \frac{1}{m} \sum_{i=1}^m (y_i \mathbf{w}^\top \phi(\mathbf{x}_i) - \gamma_m)^2 \\
&= \frac{1}{m} \sum_{i=1}^m (y_i^2 \mathbf{w}^\top \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \mathbf{w} - 2y_i \mathbf{w}^\top \phi(\mathbf{x}_i) \gamma_m + \gamma_m^2) \\
&= \frac{1}{m} \left[ \sum_{i=1}^m (\mathbf{w}^\top \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \mathbf{w}) - 2\gamma_m \sum_{i=1}^m y_i \mathbf{w}^\top \phi(\mathbf{x}_i) + \sum_{i=1}^m \gamma_m^2 \right] \\
&= \frac{1}{m} \left[ \sum_{i=1}^m (\mathbf{w}^\top \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \mathbf{w}) - 2m\gamma_m \left( \frac{1}{m} \sum_{i=1}^m y_i \mathbf{w}^\top \phi(\mathbf{x}_i) \right) + m\gamma_m^2 \right] \\
&= \frac{1}{m} (\mathbf{w}^\top \mathbf{X} \mathbf{X}^\top \mathbf{w} + (m - 2m)\gamma_m^2) \\
&= \frac{1}{m} \left( \mathbf{w}^\top \mathbf{X} \mathbf{X}^\top \mathbf{w} - \frac{1}{m} \mathbf{w}^\top \mathbf{X} \mathbf{Y} \mathbf{Y}^\top \mathbf{X}^\top \mathbf{w} \right)
\end{aligned}$$

(2) 相应优化问题为

$$\begin{aligned}
\min_{\mathbf{w}, b, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 - \mu_1 \gamma_m + \mu_2 \gamma_v + C \sum_{i=1}^m \xi_i \\
\text{s.t.} \quad & y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\
& \xi_i \geq 0, i = 1, 2, \dots, m
\end{aligned}$$

(3) 引入符号  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)^\top$ ,  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_m)^\top$ , 由此可以将优化问题改写为

$$\begin{aligned}
\min_{\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\epsilon}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{m(1-\theta)^2} (\boldsymbol{\xi}^\top \boldsymbol{\xi} + \mu \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) \\
\text{s.t.} \quad & (1-\theta) \mathbf{I} - \boldsymbol{\xi} \preceq \mathbf{Y} \mathbf{X}^\top \mathbf{w} \\
& \mathbf{Y} \mathbf{X}^\top \mathbf{w} \preceq (1+\theta) \mathbf{I} + \boldsymbol{\epsilon}
\end{aligned}$$

为了得到原问题的拉格朗日对偶问题, 引入针对两个约束的拉格朗日乘子  $\alpha, \beta$ , 得到 Lagrange 函数

$$\begin{aligned}
L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\epsilon}, \alpha, \beta) &= \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{m(1-\theta)^2} (\boldsymbol{\xi}^\top \boldsymbol{\xi} + \mu \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) + \alpha^\top [(1-\theta) \mathbf{I} - \boldsymbol{\xi} - \mathbf{Y} \mathbf{X}^\top \mathbf{w}] \\
&\quad + \beta^\top [\mathbf{Y} \mathbf{X}^\top \mathbf{w} - (1+\theta) \mathbf{I} - \boldsymbol{\epsilon}] \\
&= \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{m(1-\theta)^2} (\boldsymbol{\xi}^\top \boldsymbol{\xi} + \mu \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) - (\alpha - \beta)^\top \mathbf{Y} \mathbf{X}^\top \mathbf{w} \\
&\quad + \alpha^\top ((1-\theta) \mathbf{I} - \boldsymbol{\xi}) - \beta^\top ((1+\theta) \mathbf{I} + \boldsymbol{\epsilon})
\end{aligned}$$

对上式分别求  $\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\epsilon}$  的偏导数

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - (\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathbf{Y} \mathbf{X}^\top \\ \frac{\partial L}{\partial \boldsymbol{\xi}} &= \frac{2C}{m(1-\theta)^2} \boldsymbol{\xi} - \boldsymbol{\alpha} \\ \frac{\partial L}{\partial \boldsymbol{\epsilon}} &= \frac{2C\mu}{m(1-\theta)^2} \boldsymbol{\epsilon} - \boldsymbol{\beta}\end{aligned}$$

令以上偏导数等于 0，可以解得

$$\begin{aligned}\mathbf{w} &= \mathbf{X} \mathbf{Y} (\boldsymbol{\alpha} - \boldsymbol{\beta}) \\ \boldsymbol{\xi} &= \frac{m(1-\theta)^2 \boldsymbol{\alpha}}{2C} \\ \boldsymbol{\epsilon} &= \frac{m(1-\theta)^2 \boldsymbol{\beta}}{2C\mu}\end{aligned}$$

将解代入拉格朗日函数，即可得到对偶问题

$$\begin{aligned}\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \quad & -\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathbf{Y}^\top \mathbf{X}^\top \mathbf{X} \mathbf{Y} (\boldsymbol{\alpha} - \boldsymbol{\beta}) - \frac{m(1-\theta)^2(\mu \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta})}{4C\mu} \\ & + (1-\theta)(\boldsymbol{\alpha}^\top \mathbf{I} - (1+\theta)\boldsymbol{\beta}^\top \mathbf{I}) \\ \text{s.t.} \quad & \boldsymbol{\alpha} \succeq 0 \\ & \boldsymbol{\beta} \succeq 0.\end{aligned}$$



## 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.** (1) 不同模型的精度与 AUC 值记录 (2) 不同模型在测试数据集上的 ROC 曲线

表 1: 不同模型的精度、AUC 值

模型 指标	Logistic Regression	Decision Tree	SVM
acc. on train	0.7656	0.7498	0.7987
acc. on test	0.7642	0.6804	0.7580
AUC on test	0.8246	0.6953	0.8202

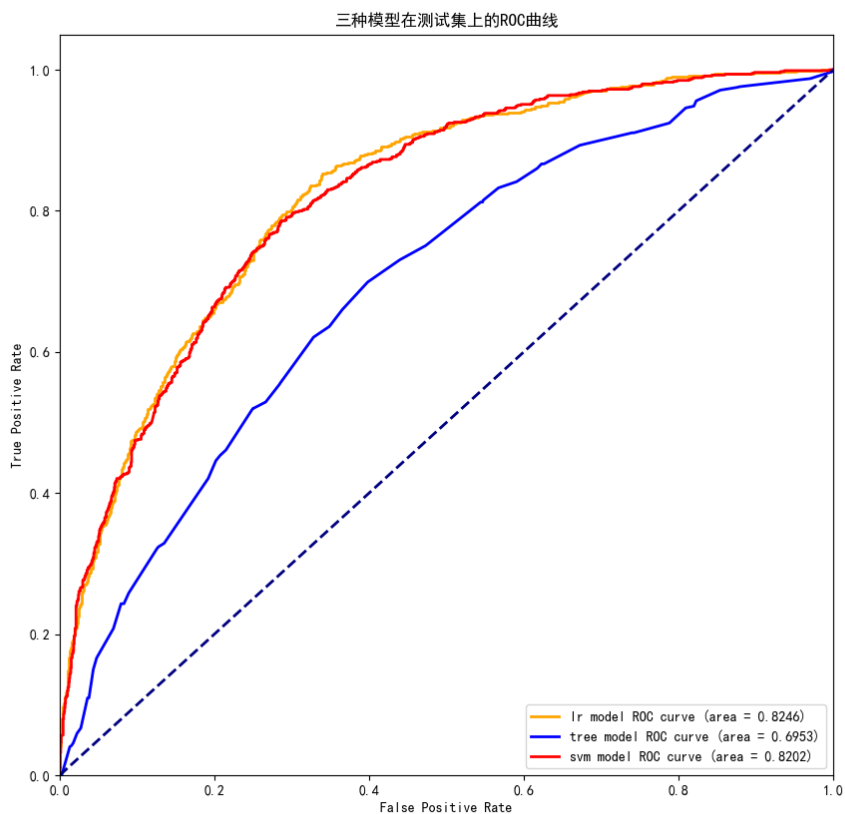
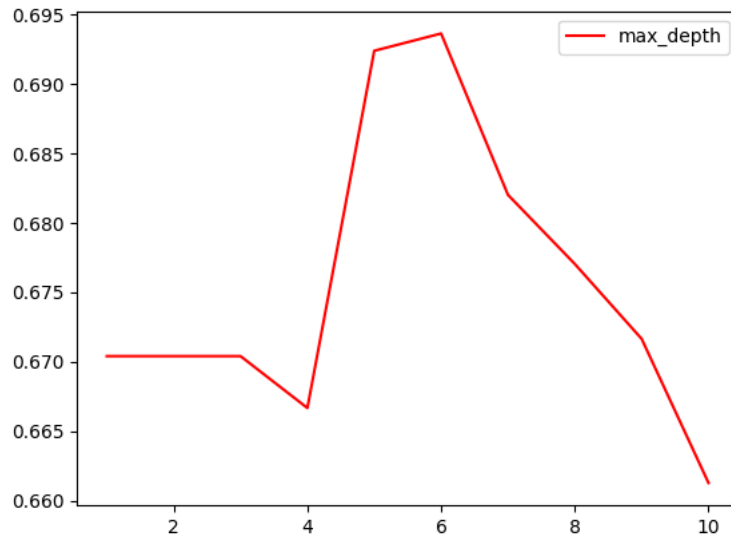


图 1: ROCs of test set

注: 在实现中`tree_clf = DecisionTreeClassifier(criterion = "gini", max_depth = 6, random_state = 30)`, 中决策树深度 `depth=6` 由以下最优剪枝参数学习曲线获得



SVM 的参数选择为:`svm_clf = svm.SVC(gamma = "scale", C = 1.0, kernel = "rbf", probability = True)`