机器学习导论 习题三

211300024, 石睿, 211300024@smail.nju.edu.cn

2023 年 5 月 1 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; 除特殊原因 (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] Representor 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:

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

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

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

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

根据第一问的结论, 该优化问题的最优解满足 $\boldsymbol{w}^\star_{\mathrm{KRR}} = \boldsymbol{X\alpha}$. 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 $\boldsymbol{w}^\star_{\mathrm{KRR}}$ 的具体形式)

Solution. 此处用于写解答 (中英文均可), 本题均以 m 作为训练集中样本数量

(1) 解:

$\because$ 若想用直和分解$\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{s}$来说明$\boldsymbol{w}^* \in \mathcal{C}(\mathcal{X})$

$\therefore$ 要对$\boldsymbol{w}^*$进行直和分解。$\boldsymbol{w}^* = \boldsymbol{v} + \boldsymbol{s} \quad (\boldsymbol{v} \in \mathcal{C}(\mathcal{X}) \quad \boldsymbol{s} \in \mathcal{C}(\mathcal{X})^\perp)$

$\therefore$ 此时可以用的子空间就只有 C(X) 啦!

$\therefore \mathcal{C}(\mathcal{X}) = span\{\phi(x_1), \cdots, \phi(x_m)\}, \boldsymbol{v} \in \mathcal{C}(\mathcal{X}), \boldsymbol{s} \in \mathcal{C}(\mathcal{X})^\perp$

$\because$ 想说明$\boldsymbol{w}^* \in \mathcal{C}(\mathcal{X})$

$\therefore$ 即说明，在对 w 的最优解直和分解之后，s=0 恒成立

$\because \min_{\boldsymbol{w}} \ J(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$的最后解为$\boldsymbol{w}^*$

【反证法】假设对于 w 的最优解，在$\mathcal{C}(\mathcal{X})$之中直和分解之后，不满足$\boldsymbol{w}^* = \boldsymbol{v}^* + \boldsymbol{s}^*$,其中$\boldsymbol{s}^* = 0$

设$\boldsymbol{w} = \boldsymbol{v}^* + s = \boldsymbol{v} + 0$,不是上述优化问题的最优解

$$\begin{cases} \boldsymbol{w}^T\phi(x_i) = (\boldsymbol{v}^*)^T\phi(x_i) \\ \|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} = (\boldsymbol{v}^*)^T(\boldsymbol{v}^*) \end{cases}$$

$\because$ 因为 w 不是优化问题最优解，应有$J(\boldsymbol{w}) - J(\boldsymbol{w}^*) > 0$

$$[1]J(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left((\boldsymbol{v}^*)^T\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}(\boldsymbol{v}^*)^T(\boldsymbol{v}^*)$$

$$J(\boldsymbol{w}^*) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(((\boldsymbol{v}^*)^T + (\boldsymbol{s}^*)^T)\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}\left((\boldsymbol{v}^*)^T(\boldsymbol{v}^*) + (\boldsymbol{s}^*)^T(\boldsymbol{s}^*)\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left((\boldsymbol{v}^*)^T\phi(\boldsymbol{\xi}_i) + (\boldsymbol{s}^*)^T\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}\left((\boldsymbol{v}^*)^T(\boldsymbol{v}^*) + (\boldsymbol{s}^*)^T(\boldsymbol{s}^*)\right)$$

$$\because \boldsymbol{s}^* \in \mathcal{C}(\mathcal{X})^\perp, \phi(x_i) \in \mathcal{C}(\mathcal{X})$$

$$\therefore (\boldsymbol{s}^*)^T\phi(x_i) = 0$$

$$[2]J(\boldsymbol{w}^*) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left((\boldsymbol{v}^*)^T\phi(\boldsymbol{\xi}_i), y_i\right) + \frac{\lambda}{2}\left((\boldsymbol{v}^*)^T(\boldsymbol{v}^*) + (\boldsymbol{s}^*)^T(\boldsymbol{s}^*)\right)$$

$$\therefore [1]-[2] = J(\boldsymbol{w}) - J(\boldsymbol{w}^*) = -\frac{\lambda}{2}(\boldsymbol{s}^*)^T(\boldsymbol{s}^*)$$

$$\because \boldsymbol{s}^* \neq 0$$

$\therefore J(\boldsymbol{w}) - J(\boldsymbol{w}^*) < 0,$ 和 w 不是最优解相矛盾！假设不成立

（2）解：

$\because \|\boldsymbol{w}\|^2, (\boldsymbol{w}^T\phi(x_i) - y_i)^2,$ 分别是关于 w 的凸函数，仿射 + 平方的复合函数。均为凸函数

$\therefore \min\limits_{\boldsymbol{w}} F(\boldsymbol{w})$是凸优化问题

$\therefore \boldsymbol{w}^*$是最优解的必要条件的为$\nabla F(\boldsymbol{w}^*) = 0$

$$F(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m}\left(\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i) - y_i\right)^2$$

$$= \lambda\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{m}\left((\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i))(\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i))^T - 2\boldsymbol{w}^T\phi(x_i)y_i + y_i^2\right)^2$$

$$= \lambda\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{m}\left(\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i)\phi(\boldsymbol{\xi}_i)^T\boldsymbol{w} - 2\boldsymbol{w}^T\phi(x_i)y_i + y_i^2\right)^2$$

$$\frac{\partial F(\boldsymbol{w})}{\partial \boldsymbol{w}} = \nabla F(\boldsymbol{w}) = 2\lambda\boldsymbol{w} + \sum_{i=1}^{m}\left(2\phi(\boldsymbol{\xi}_i)\phi(\boldsymbol{\xi}_i)^T\boldsymbol{w} - 2\phi(x_i)y_i\right)^2$$

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

令$\dfrac{\partial F(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$

$$\therefore (\lambda E + \boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{w} = \boldsymbol{X}\boldsymbol{y}$$

$$\therefore \boldsymbol{w} = (\lambda E + \boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{y} = \boldsymbol{X}(\lambda E + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{y}$$

$$\therefore \alpha = (\lambda E + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{y}$$

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

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

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

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

(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 $\boldsymbol{\xi}$ 并非 h_S 的支持向量, 则移除该样本不会改变 SVM 模型, 即 $h_{S\setminus\{\boldsymbol{\xi}\}}=h_S$. 这一性质在分析误差时有关键作用, 考虑如下问题: 如果 $\boldsymbol{\xi}$ 不是 h_S 的支持向量, $h_{S\setminus\{\boldsymbol{\xi}\}}$ 会将 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:

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

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) 解:

$$\because\hat{R}_{\mathrm{LOO}}(\mathcal{A})=\frac{1}{m}\sum_{i=1}^{m}1_{h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i}=\frac{1}{m}\sum_{i=1}^{m}X_i$$

其中 X 为新定义的随机变量, $X_i=1_{h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i}=I(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)=\begin{cases}1 & h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i\\ 0 & \text{otherwise}\end{cases}$

$$\therefore\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})]=\frac{1}{m}\sum_{i=1}^{m}E(X_i)$$

$\therefore$ 以下只需要求$E(X_i)$

$\because$ 由指示函数 I 的定义可知

$[1]E(X_i)=E_{S\sim\mathcal{D}^m}(I(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i))=P_{S\sim\mathcal{D}^m}(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)$

$\because$ 由泛化误差的定义

$$[2]R\left(h_{S'}\right)=\frac{1}{m}\sum_{i=1,(x,y)\sim D}^{m} I(h_{S'}(\boldsymbol{\xi}_i)\neq y_i)$$
$$=E(X)$$
$$=P_{(x,y)\sim D}(h_{S'}(\boldsymbol{\xi}_i)\neq y_i)$$

由 [1][2] 两式，以下只需要证明 [***] 式

$$[***]\quad P_{S\sim D^m}(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)=E_{S'\sim D^{m-1}}\left(P_{S\sim\mathcal{D}^m}(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)\right)$$

$\because S'=S\setminus\{\boldsymbol{\xi}_i\}$,其中$(x_i,y_i)\sim D^m$

$\because$ 对数据集的分布 D 进行采样是独立同分布的

$\therefore (x_i,y_i)\sim D$

$$\therefore P_{S\sim D^m,(x_i,y_i)\sim D^m}(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)=P_{S'\sim D^{m-1},(x_i,y_i)\sim D}(h_{S'}(\boldsymbol{\xi}_i)\neq y_i)$$
$$=P_{S'\sim D^{m-1},(x,y)\sim D}(h_{S'}(\boldsymbol{\xi})\neq y)\quad \text{(tip: 为了统一符号，全部以 x 和 y 来代替)}$$
$$=\sum_{(x,y)\sim D} P_{S'\sim D^{m-1}}(h_{S'}(\boldsymbol{\xi})\neq y|(x,y)\sim D)\cdot P((x,y)\sim D)\quad \text{(tip: 全概率公式)}$$
$$=E_{S'\sim D^{m-1},(x,y)\sim D}\left(P(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)\right)$$
$$=E_{S'\sim D^{m-1}}\left(P_{(x,y)\sim D}(h_{S\setminus\{\boldsymbol{\xi}_i\}}(\boldsymbol{\xi}_i)\neq y_i)\right)$$

(2) 解：

[1]$h_{S\setminus\{\boldsymbol{\xi}\}}$会把 x 正确分类。由题目提示，此数据集线性可分

$\therefore$ 一定存在$\boldsymbol{w}^T\boldsymbol{\xi}+b$使 S 可以被正确分类

$\because$ x 不是支持向量，所以$h_{S\setminus\{\boldsymbol{\xi}\}}=h_S$

注意到h_S可以把 x 争取分类

$\therefore h_{S\setminus\{\boldsymbol{\xi}\}}$也可以把 x 正确分类呢!

[2]逆否命题：若$h_{S\setminus\{\boldsymbol{\xi}\}}$没有把 x 争取分类，则 x 是$h_S$的支持向量

(3) 解：

$\because \mathbb{E}_{S'\sim\mathcal{D}^{m+1}}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] = \mathbb{E}_{S\sim\mathcal{D}^m}\left[R\left(h_S\right)\right]$

也即，可以通过留一法的测试误差来无偏估计泛化误差

由留一法的测试误差定义$\hat{R}_{\mathrm{LOO}}(\mathcal{A}) = \dfrac{1}{m+1}\displaystyle\sum_{i=1}^{m} I(h_{S\setminus\{x_i\}} \neq y_i)$

$\therefore E_{S'\sim D^{m+1}}(\hat{R}_{\mathrm{LOO}}(\mathcal{A})) = E_{S'\sim D^{m+1}}\left[\dfrac{\sum_{i=1}^{m+1} E(I(h_{S\setminus\{x_i\}} \neq y_i))}{m+1}\right]$

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

只需证明$\displaystyle\sum_{i=1}^{m+1} E_{(x_i,y_i)\in S'}(I(h_{S\setminus\{x_i\}} \neq y_i)) \leq N_{SV}(S)$

以下对S'中的(x_i, y_i)是否为支持向量进行讨论

$[situation1]$　(x_i, y_i)不是h_S的支持向量

$\therefore$ 由第二问的结论　$I(h_{S\setminus\{x_i\}} \neq y_i) = 0$

$\therefore (x_i, y_i)$ $N_{SV}(S)$没有贡献，同时也对 E 没有贡献

$[situation2]$　(x_i, y_i)是h_S的支持向量

由 (2) 的逆否命题，既可以得知

可能 x 是h_S的支持向量，但$h_{S\setminus\{x\}}$仍然可能在 x 处正确分类

$\therefore I(h_{S\setminus\{x_i\}} \neq y_i) = 0/1$

$\therefore x$对$N_{SV}(S)$有贡献的时候，但仍然 x 可能对 E 没有贡献

$[situation1+2]$　综上，可以得出$\displaystyle\sum_{i=1}^{m+1} E_{(x_i,y_i)\in S'}(I(h_{S\setminus\{x_i\}} \neq y_i)) \leq N_{SV}(S)$

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 $\mathcal{S} = \{(\boldsymbol{\xi}_1, y_1), \cdots, (\boldsymbol{\xi}_m, y_m)\}$, $\phi : \mathcal{X} \to \mathbb{H}$ 为映射函数, 我们记 $\boldsymbol{X} = [\phi(\boldsymbol{\xi}_1), \cdots, \phi(\boldsymbol{\xi}_m)]$ 为映射后的数据矩阵, $\boldsymbol{y}^T = [y_1, \cdots, y_m]$ 为标签向量, $\boldsymbol{Y}$ 是对角元素为 $y_1, \cdots, y_m$ 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:

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

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, 我们希望在其基础上进行改进: 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 $(\boldsymbol{\xi}_i, y_i)$ 的间隔相较于均值的偏移为 $|y_i \boldsymbol{w}^T \phi(\boldsymbol{\xi}_i) - 1|$. 此时仅需最小化间隔方差, 相应的优化问题为

$$\begin{aligned} \min_{\boldsymbol{w}, \xi_i, \epsilon_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\left(\xi_i^2 + \epsilon_i^2\right) \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi\left(\boldsymbol{\xi}_i\right) \geq 1 - \xi_i, y_i \boldsymbol{w}^\top \phi\left(\boldsymbol{\xi}_i\right) \leq 1 + \epsilon_i, \forall i. \end{aligned}$$

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:

$$\begin{aligned} \min_{\boldsymbol{w}, \xi_i, \epsilon_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\frac{\xi_i^2 + \mu\epsilon_i^2}{(1-\theta)^2} \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi\left(\boldsymbol{\xi}_i\right) \geq 1 - \theta - \xi_i \\ & y_i \boldsymbol{w}^\top \phi\left(\boldsymbol{\xi}_i\right) \leq 1 + \theta + \epsilon_i, \forall i. \end{aligned}$$

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) 解:

$$
\begin{aligned}
\gamma_m &= \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{w}^T \phi(\boldsymbol{\xi}_i) \\
&= \frac{1}{m}\Big(\boldsymbol{w}^T\phi(x_i), \quad \cdots \quad, \boldsymbol{w}^T\phi(x_m)\Big)\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \\
&= \frac{1}{m}\boldsymbol{w}^T\boldsymbol{X}y \\
&= \frac{1}{m}y^T\boldsymbol{X}^T\boldsymbol{w}
\end{aligned}
$$

$$
\begin{aligned}
\gamma_v &= \frac{1}{m}\sum_{i=1}^{m}(y_i\boldsymbol{w}^T\phi(\boldsymbol{\xi}_i) - \gamma_m)^2 \\
&= \frac{1}{m}\left[\sum_{i=1}^{m}(y_i\boldsymbol{w}^T\phi(x_i))^2 - 2\gamma_m\sum_{i=1}^{m}y_i\boldsymbol{w}^T\phi(x_i) + \gamma_m^2 m\right] \\
&= \frac{1}{m}\left[\boldsymbol{w}^T\boldsymbol{X}y^Ty\boldsymbol{X}^T\boldsymbol{w} - 2\frac{1}{m}\boldsymbol{w}^T\boldsymbol{X}yy^T\boldsymbol{X}^T\boldsymbol{w} + \frac{1}{m}\boldsymbol{w}^T\boldsymbol{X}yy^T\boldsymbol{X}^T\boldsymbol{w}\right] \\
&= \frac{1-m}{m^2}\boldsymbol{w}^T\boldsymbol{X}y^Ty\boldsymbol{X}^T\boldsymbol{w}
\end{aligned}
$$

(2) 解:

$$
\text{类比} \quad \min_{w,b} \quad \frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{m} max(0, 1 - y_i(\boldsymbol{w}^T x_i + b))
$$

$\therefore$ 本问题可以写成如下优化形式，引入了松弛变量ξ_i

$$
\begin{aligned}
&\min_{\boldsymbol{w},b} \quad \frac{1}{2}\left\|w\right\|^2 - \frac{\mu_1}{m}\gamma_m + \frac{\mu_2}{m}\gamma_v + C\sum_{i=1}^{m}\xi_i \\
&\text{s.t.} \quad y_i(\boldsymbol{w}^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0
\end{aligned}
$$

(3) 解：

$$\text{本小问引入拉格朗日乘子}\boldsymbol{\alpha}=\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_m\end{pmatrix}\geq 0 \quad \boldsymbol{\beta}=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_m\end{pmatrix}\geq 0$$

$$\begin{aligned}F&=\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{C}{m}\sum_{i=1}^{m}\frac{\xi_i^2+\mu\epsilon_i^2}{(1-\theta)^2}\\&=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\frac{C}{m}\frac{\xi^T\xi}{(1-\theta)^2}+\frac{C}{m}\frac{\mu\epsilon^T\epsilon}{(1-\theta)^2}\end{aligned}$$

[$**step1**$]对 F(x) 引入拉格朗日乘子来构造拉格朗日函数

$$\begin{aligned}L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta})&=F+\sum_{i=1}^{m}\alpha_i\cdot(1-\theta-\xi_i-y_i\boldsymbol{w}^T\phi(x_i))+\sum_{i=1}^{m}\beta_i\cdot(y_i-\boldsymbol{w}^T\phi(x_i)-1-\theta-\epsilon_i)\\&=F-\boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{\beta}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{\beta}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}\\&=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\frac{C}{m}\frac{\xi^T\xi}{(1-\theta)^2}+\frac{C}{m}\frac{\mu\epsilon^T\epsilon}{(1-\theta)^2}-\boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{\beta}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}+\boldsymbol{\beta}^T\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}\end{aligned}$$

[$**step2**$]由 KKT 条件（或可以由凸函数最优解的必要条件）进行化简

$\because$ 拉格朗日对偶函数为$\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})=\inf\limits_{\boldsymbol{w},\boldsymbol{\epsilon},\boldsymbol{\xi}}L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta})$

$\because L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta})$是关于$\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon}$的凸优化问题，且满足 slater 条件, 故满足强对偶性

$\therefore$ 此时凸优化问题的 KKT 条件是充分必要的

$\therefore \boldsymbol{w}^*,\boldsymbol{\xi}^*,\boldsymbol{\epsilon}^*$可以让$\dfrac{\partial L}{\partial\boldsymbol{w}^*}=0,\dfrac{\partial L}{\partial\boldsymbol{\xi}^*}=0,\dfrac{\partial L}{\partial\boldsymbol{\epsilon}^*}=0$

$$\frac{\partial L}{\partial\boldsymbol{w}}=\boldsymbol{w}-(\boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{X}^T)^T+(\boldsymbol{\beta}\boldsymbol{Y}\boldsymbol{X}^T)^T=\boldsymbol{w}-\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{\alpha}+\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{\beta}=0$$

[1] $\therefore \boldsymbol{w}^*=\boldsymbol{X}\boldsymbol{Y}^T(\boldsymbol{\alpha}-\boldsymbol{\beta})$

$$\frac{\partial L}{\partial\boldsymbol{\xi}}=\frac{C}{m}\cdot 2\cdot\boldsymbol{\xi}\cdot\frac{1}{(1-\theta)^2}-\boldsymbol{\alpha}=0$$

[2] $\therefore \boldsymbol{\xi}^*=\dfrac{m(1-\theta)^2\boldsymbol{\alpha}}{2C}$

$$\frac{\partial L}{\partial\boldsymbol{\epsilon}}=\frac{C}{m}\cdot 2\cdot\boldsymbol{\epsilon}\cdot\frac{\mu}{(1-\theta)^2}-\boldsymbol{\beta}=0$$

[3] $\therefore \boldsymbol{\epsilon}^*=\dfrac{m(1-\theta)^2\boldsymbol{\beta}}{2C\mu}$

[$**step3**$]把 [1][2][3] 的结果带入到$L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta})$

$$\begin{aligned}\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})&=\inf_{\boldsymbol{w},\boldsymbol{\epsilon},\boldsymbol{\xi}}L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\boldsymbol{\alpha},\boldsymbol{\beta})\\&=-\frac{1}{2}(\boldsymbol{\alpha}^T-\boldsymbol{\beta}^T)\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{Y}(\boldsymbol{\alpha}-\boldsymbol{\beta})\\&-\frac{m(1-\theta)(\mu\boldsymbol{\alpha}^T\boldsymbol{\alpha}+\boldsymbol{\beta}^T\boldsymbol{\beta}))}{4C\mu}+(1-\theta)(\boldsymbol{\alpha}^TE-(1-\theta)\boldsymbol{\beta}^TE)\end{aligned}$$

[$******$] $\therefore$ 拉格朗日对偶问题如下

$$\begin{aligned}&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\quad\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})\\&\text{s.t.}\quad\boldsymbol{\alpha}\geq 0,\quad\boldsymbol{\beta}\geq 0\end{aligned}$$

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.

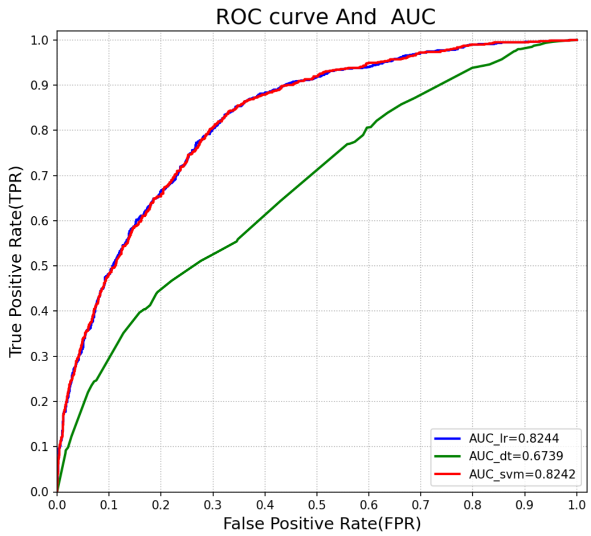
Solution. 此处用于写解答 (中英文均可)

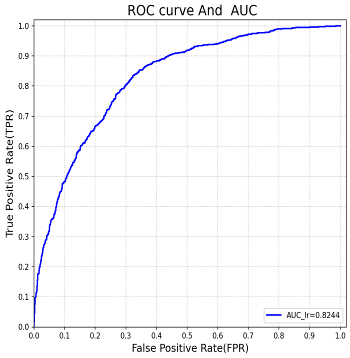
(1) 不同模型的精度与 AUC 值记录

表 1: 不同模型的精度、AUC 值

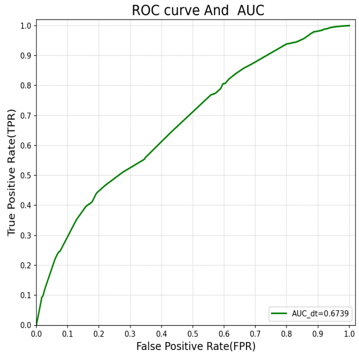
模型 指标	Logistic Regression	Decision Tree	SVM
acc. on train	76.58%	71.88%	76.85%
acc. on test	76.38%	69.49%	76.26%
AUC on test	0.8244	0.6739	0.8242

(2) 不同模型在测试数据集上的 ROC 曲线

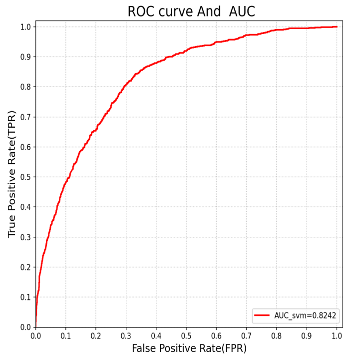




[对数几率]



[决策树]



[支持向量机]