机器学习理论研究导引
作业一

庄镇华 502022370071

2023 年 3 月 20 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2023/03/30 23:59:59**, 截止时间后不再接收作业, 本次作业记零分;

(2) 作业提交方式: 使用此 LaTex 模板书写解答, 只需提交编译生成的 pdf 文件, 将 pdf 文件提交至南大网盘:
https://box.nju.edu.cn/u/d/710f38cc21884d798b12/

(3) pdf 文件命名方式: 学号-姓名-作业号-v 版本号, 例 MG1900000-张三-1-v1; 如果需要更改已提交的解答, 请在截止时间之前提交新版本的解答, 并将版本号加一;

(4) 未按照要求提交作业, 或 **pdf 命名方式不正确**, 将会被扣除部分作业分数.

1 [25pts] Exponentially Concave Functions

指数凹函数是一类在 prediction with expert advice 中广泛研究的函数. 我们称一个函数 $f(x)$ 是指数凹的, 若存在常数 $\eta > 0$, 使得 $\mathrm{e}^{-\eta f(x)}$ 是凹函数.

(1) **[15pts]** 请证明指数凹函数一定是凸函数.

(2) **[10pts]** 凸函数是否一定为指数凹函数? 请给出证明.

Solution.

(1) 证明:

首先证明定义在凸集 Ψ 上的函数 $f(x)$ 是凸函数的充要条件为 $\forall x, y \in \Psi$, 均满足 $f(x) \geq f(y) + \nabla f(y)^T (x - y)$。

必要性: 当 $f(x)$ 为凸函数时, $\forall x, y \in \Psi$, 均满足 $f\left(\theta x + (1-\theta)y\right) \leqslant \theta f\left(x\right) + (1-\theta) f\left(y\right)$, $(\forall 0 \leqslant \theta \leqslant 1)$, 两端同时除以 θ, 得到 $f(x) \geq f(y) + \frac{f(y+\theta(x-y))-f(y)}{\theta(x-y)}(x-y)$, 当 $\theta(x-y)$ 趋于 0 时, 取极限, 可得 $f(x) \geq f(y) + \nabla f(y)^T (x - y)$。

充分性: 选择任意 $x \neq y, 0 \leq \theta \leq 1$, 令 $z = \theta x + (1-\theta)y$, 可得 $f(x) \geqslant f(z) + \nabla f(z)(x-z)$, $f(y) \geqslant f(z) + \nabla f(z)(y-z)$, 将前式乘以 θ, 后式乘以 $1-\theta$, 相加可得 $\theta f(x) + (1-\theta) f(y) \geqslant f(z) = f(\theta x + (1-\theta)y)$, 即 $f(x)$ 为凸函数。

同理可证 $f(x)$ 是凹函数的充要条件为 $\forall x, y \in \Psi$, 均满足 $f(x) \leq f(y) + \nabla f(y)^T (x - y)$。

接下来证明指数凹函数一定是凸函数。因为 $\mathrm{e}^{-\eta f(x)}$ 是凹函数, 所以 $\forall x, y \in \Psi$, 均满足 $\mathrm{e}^{-\eta f(x)} \leq \mathrm{e}^{-\eta f(y)} + \nabla \mathrm{e}^{-\eta f(y)}(x-y) = \mathrm{e}^{-\eta f(y)} - \eta \nabla f(y) \mathrm{e}^{-\eta f(y)}(x-y)$, 两边同时乘以 $\mathrm{e}^{\eta f(y)}$, 可得 $\mathrm{e}^{\eta f(y) - \eta f(x)} \leq 1 - \eta \nabla f(y)(x-y)$, 又因为 $\mathrm{e}^{\eta f(y) - \eta f(x)} \geq \eta f(y) - \eta f(x) + 1$, 所以 $1 - \eta \nabla f(y)(x-y) \geq \eta f(y) - \eta f(x) + 1$, 整理可得 $f(x) \geq f(y) + \nabla f(y)^T (x-y)$, 即 $f(x)$ 是凸函数, 得证。

(2) 凸函数不一定为指数凹函数。证明如下:

易知仿射函数 $f(x) = x$ 是凸函数 $(x \geq y + 1(x-y))$, 假设存在 $\eta > 0$, 使得 $\mathrm{e}^{-\eta x}$ 是凹函数, 即 $\forall x, y \in \Psi$, 均满足 $\mathrm{e}^{-\eta x} \leq \mathrm{e}^{-\eta y} - \eta \mathrm{e}^{-\eta y}(x-y)$, 两边同时乘以 $\mathrm{e}^{\eta y}$, 即 $\mathrm{e}^{\eta y - \eta x} \leq 1 - \eta(x-y)$, 即 $\mathrm{e}^{\eta(y-x)} \leq 1 + \eta(y-x)$, 因为上式对于 $\forall x, y \in \Psi$ 均成立, 所以 $\eta = 0$, 这与 $\eta > 0$ 矛盾, 因此不存在 $\eta > 0$, 使得 $\mathrm{e}^{-\eta x}$ 是凹函数, 即凸函数 $f(x) = x$ 不是指数凹函数, 因此凸函数不一定为指数凹函数, 得证。

2 [25pts] Dual Problem

支持向量机的思想是最大化最小间隔, 而研究 [Gao and Zhou, 2013] 表明优化间隔的分布可以取得更好的泛化性能. 基于此, 最优间隔分布机 (Optimal margin Distribution Machine) 考虑同时最小化间隔的均值和最大化间隔的方差.

给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$, 最优间隔分布机的优化问题为:

$$\begin{aligned} \min_{\boldsymbol{w},\xi_i,\epsilon_i} & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\frac{\xi_i^2 + \mu\epsilon_i^2}{(1-\theta)^2}, \\ \text{s.t. } & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1-\theta-\xi_i \\ & y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \leq 1+\theta+\epsilon_i, \forall i \end{aligned} \tag{2.1}$$

其中 $\lambda > 0$ 是正则项系数, μ 表示内侧样本损失的相对权重, θ 控制支持向量的数目. 试推导上述问题的对偶问题.

相关文献:

- Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. **Artificial Intelligence**, 2013, 203: 1-18.

Solution.

推导对偶问题

易知 ξ_i 和 ϵ_i 是度量实例损失的松弛变量, 定义 $\boldsymbol{\xi}^\top = [\xi_1, \cdots, \xi_m]$, 同理可定义 $\boldsymbol{\epsilon}$, $\boldsymbol{Y}$, $\boldsymbol{X}$。

为式*2.1*中的两组不等式约束引入拉格朗日乘子 $\boldsymbol{\alpha} \geq \boldsymbol{0}$ 和 $\boldsymbol{\beta} \geq \boldsymbol{0}$, 定义 $\boldsymbol{e}$ 为全 *1* 向量, 可知拉格朗日函数为

$$\mathcal{L} = \frac{||\boldsymbol{w}||^2}{2} + \frac{\lambda\boldsymbol{\xi}^\top\boldsymbol{\xi}}{m(1-\theta)^2} + \frac{\lambda\mu\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}}{m(1-\theta)^2} - (\boldsymbol{\alpha}-\boldsymbol{\beta})^\top\boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{w} + \boldsymbol{\alpha}^\top((1-\theta)\boldsymbol{e}-\boldsymbol{\xi}) - \boldsymbol{\beta}^\top((1+\theta)\boldsymbol{e}+\boldsymbol{\epsilon})$$

令 $\mathcal{L}$ 对 $\boldsymbol{w}$, $\boldsymbol{\xi}$, $\boldsymbol{\epsilon}$ 的偏导等于 *0*, 我们得到

$$\boldsymbol{w} = \boldsymbol{X}\boldsymbol{Y}(\boldsymbol{\alpha}-\boldsymbol{\beta}), \boldsymbol{\xi} = \frac{m(1-\theta)^2\boldsymbol{\alpha}}{2\lambda}, \boldsymbol{\epsilon} = \frac{m(1-\theta)^2\boldsymbol{\beta}}{2\lambda\mu}$$

带入, 消除 $\boldsymbol{w}$, 得到主问题的对偶问题

$$\mathcal{L} = -\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^\top\boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{Y}(\boldsymbol{\alpha}-\boldsymbol{\beta}) - \frac{m(1-\theta)^2(\mu\boldsymbol{\alpha}^T\boldsymbol{\alpha}+\boldsymbol{\beta}^T\boldsymbol{\beta})}{4\lambda\mu} + (1-\theta)\boldsymbol{\alpha}^T\boldsymbol{e} - (1+\theta)\boldsymbol{\beta}^T\boldsymbol{e}$$

令 $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{Y}$, $\boldsymbol{\gamma}^T = [\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T]$, 那么 $\boldsymbol{\alpha} = [\boldsymbol{I}, \boldsymbol{0}]\boldsymbol{\gamma}$, $\boldsymbol{\beta} = [\boldsymbol{0}, \boldsymbol{I}]\boldsymbol{\gamma}$, $\boldsymbol{\alpha}-\boldsymbol{\beta} = [\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{\gamma}$, 则上式可化简为

$$\min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}^T}{2}\begin{bmatrix} \boldsymbol{Q}+\frac{m(1-\theta)^2}{2\lambda}\boldsymbol{I} & -\boldsymbol{Q} \\ -\boldsymbol{Q} & \boldsymbol{Q}+\frac{m(1-\theta)^2}{2\lambda\mu}\boldsymbol{I} \end{bmatrix}\boldsymbol{\gamma} + \begin{bmatrix} (\theta-1)\boldsymbol{e} \\ (\theta+1)\boldsymbol{e} \end{bmatrix}^T\boldsymbol{\gamma}, \; s.t. \boldsymbol{\gamma} \geq \boldsymbol{0}$$

预测标签

对于预测，可从最优 $\boldsymbol{\gamma}^\star$ 获得最优系数 $\boldsymbol{w}$

$$\boldsymbol{w} = \boldsymbol{X}\boldsymbol{Y}(\boldsymbol{\alpha} - \boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{Y}[\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{\gamma}^\star = \boldsymbol{X}\boldsymbol{v}$$

其中 $\boldsymbol{v} = \boldsymbol{Y}[\boldsymbol{I}, -\boldsymbol{I}]\boldsymbol{\gamma}^\star$，因此对于测试样例 $\boldsymbol{z}$，其标签为 $sign\left(\boldsymbol{w}^T\phi\left(\boldsymbol{z}\right)\right) = sign\left(\sum_{i=1}^{m} v_i\kappa\left(\boldsymbol{x}_i, \boldsymbol{z}\right)\right)$。

参考文献:

- *Teng Zhang and Zhi-Hua Zhou. Optimal Margin Distribution Machine.* ***IEEE Transactions on Knowledge and Data Engineering****, 2020, 32: 1143-1156.*

3 [25pts] PAC Learning for Finite Hypothesis Sets

课件中通过构造一个精巧的算法证明了布尔合取式概念类的可学习性. 事实上, 对于可分的有限假设空间, 简单的 ERM 算法也可以导出 PAC 可学习性. 请证明:

令 $\mathcal{H}$ 为可分的有限假设空间, D 为包含 m 个从 $\mathcal{D}$ 独立同分布采样所得的样本构成的训练集, 学习算法 $\mathfrak{L}$ 基于训练集 D 返回与训练集一致的假设 h_D, 对于任意 $c \in \mathcal{H}$, $0 < \epsilon, \delta < 1$, 如果有 $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$, 则

$$P\left(E(h_D) \leq \epsilon\right) \geq 1 - \delta, \tag{3.1}$$

即 $E(h) \leq \epsilon$ 以至少 $1-\delta$ 的概率成立.

提示: 注意到 h_D 必然满足 $\widehat{E}_D(h_D) = 0$.

Proof.

证明：首先估计泛化误差大于 ϵ 但在训练集上表现完美的假设出现的概率。假设 h 的泛化误差大于 ϵ，那么对分布 $\mathcal{D}$ 上随机采样得到的任何样本 (x, y)，有

$$P(h(x) = y) = 1 - P(h(x) \neq y) = 1 - E(h) < 1 - \epsilon$$

因为 D 包含 m 个从 $\mathcal{D}$ 独立同分布采样而得的样本，所以 h 在训练集 D 上表现完美的概率为

$$P((h(x_1) = y_1) \wedge \cdots \wedge (h(x_m) = y_m)) = (1 - P(h(x) \neq y))^m < (1-\epsilon)^m$$

泛化误差大于 ϵ，且在训练集上表现完美的所有假设出现概率之和为

$$P(h \in \mathcal{H}{:}E(h) > \epsilon \wedge \hat{E}(h) = 0) < |\mathcal{H}|(1-\epsilon)^m < |\mathcal{H}|e^{-m\epsilon}$$

只需保证其不大于 δ，即

$$|\mathcal{H}|e^{-m\epsilon} \leq \delta$$

可得

$$m \geq \frac{1}{\epsilon}(ln|\mathcal{H}| + ln\frac{1}{\delta})$$

所以，有限假设空间 $\mathcal{H}$ 都是 PAC 可学习的，所需的样本数目如上式所示，也即如果有 $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$，则 $E(h) \leq \epsilon$ 以至少 $1-\delta$ 的概率成立。

4 [25pts] PAC Learning for Infinite Hypothesis Sets

课件中已经证明了轴平行矩形的假设空间是可学习的. 这启发我们, 无限假设空间也可能是可学习的. 本题给出另一个可学习的无限假设空间的简单的例子.

令 $\mathcal{H} = \{h_r : \mathbb{R}^2 \to \{-1, +1\} \mid h_r(\boldsymbol{x}) = \mathbb{I}(x^2 + y^2 \leqslant r^2), r \in \mathbb{R}\}$ 表示以原点为圆心的圆构成的假设空间. 假设目标概念 $c \in \mathcal{H}$. 试证明这个无限假设空间 $\mathcal{H}$ 是 PAC 可学的.

Proof.

证明: 考虑学习算法 $\mathcal{L}$: 对于训练集 D, $\mathcal{L}$ 输出一个包含 D 中所有正例的最小圆 $\mathcal{R}^D$, 显然, $\mathcal{R}^D$ 中的点一定包含在目标概念 c 对应的圆 R 中, 因此 $\mathcal{R}^D$ 不会将负例误判为正例, 它犯错误的区域都包含在 R 中。

令 $P(R)$ 表示 R 区域的概率质量, 即按照分布 $\mathcal{D}$ 随机生成的点落在区域 R 中的概率。由于学习算法 $\mathcal{L}$ 的错误仅可能出现在 R 内的点上, 不妨设 $P(R) > \epsilon$, 否则无论输入什么训练集 D, R^D 的错误率都不会超过 ϵ。

因为 $P(R) > \epsilon$, 并且假设空间中所有圆的圆心都在原点, 因此我们可以定义 *1* 个圆环区域 $r = R - R'$, 其中 R' 的圆心在原点上并且 $P(R) - P(R') = \epsilon$, 也即 $P(r) = \epsilon$。

由于 R^D 位于 R 内部, 因此若 R^D 与 r 相交, 那么 R^D 的错误区域被 r 覆盖, 其泛化误差 $E(R^D) \leq \epsilon$。

于是, 若泛化误差 $E(R^D) > \epsilon$, 则 R^D 必然与 r 不相交。训练集 D 中的每个样本是从分布 $\mathcal{D}$ 随机采样得到的, 其出现在 r 中的概率为 ϵ。设 D 中包含 m 个样本, 则有

$$P_{D\sim\mathcal{D}^m}(E(R^D) > \epsilon) \leq P_{D\sim\mathcal{D}^m}(R^D \cap r = \varnothing) \leq (1-\epsilon)^m \leq e^{-m\epsilon}$$

令 $e^{-m\epsilon} \leq \delta$, 即可确保

$$P_{D\sim\mathcal{D}^m}(E(R^D) \leq \epsilon) = 1 - P_{D\sim\mathcal{D}^m}(E(R^D) > \epsilon) \geq 1 - \delta$$

于是可求解得到

$$m \geq \frac{1}{\epsilon} ln \frac{1}{\delta}$$

上面构造的学习算法仅需上式所示的样本数就能以至少 $1-\delta$ 的概率得到满足 $E(R^D) \leq \epsilon$ 的假设 R^D, 且该学习算法所涉及的 $\mathbb{R}^2$ 中以原点为圆心的圆的计算时间为常数, 于是可知该无限假设空间 $\mathcal{H}$ 是 *PAC* 可学的。