机器学习理论研究导引
作业一

魏沐昊 **********

2024 年 4 月 26 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2024/04/03 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/ae7c31c933584e03a095/

(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;

(5) 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] Kernel Functions

(1) **[10 pts]** 考虑 $\mathbb{R}^N$ 上的函数 $\kappa(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^\top\boldsymbol{y}+c)^d$, 其中 c 是任意实数, d,N 是任意正整数. 试分析函数 κ 何时是核函数, 何时不是核函数, 并说明理由.

(2) **[15 pts]** 当上一小问中的函数是核函数时, 考虑 $d=2$ 的情况, 此时 κ 将 N 维数据映射到了什么空间中? 具体的映射函数是什么? 更一般的, 对 d 不加限制时, κ 将 N 维数据映射到了什么空间中? (本小问的最后一问可以只写结果)

Solution.

(1). 对于该多项式函数，根据核函数的 *Mercer* 条件，当该函数满足如下条件*: 1.* 该函数是对称函数*; 2.* 由该函数构成的矩阵 $K_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{y}_j)$ 是半正定矩阵，即对于任意一组数据 $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_m\}$ 均有 $\alpha^\top K\alpha\succeq 0$ 时, 该函数为核函数。考虑一个多项式核 $\hat{\kappa}(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^\top\boldsymbol{y})^d$, 该多项式显然满足 *Mercer* 条件, 其证明如下*:*

$$\alpha^\top\hat{K}\alpha=\sum_{i=1}^m\sum_{j=1}^m\alpha_i\alpha_j(\boldsymbol{x}_i)^d(\boldsymbol{x}_j)^d \tag{1.1}$$

$$=\|\sum_{i=1}^m\alpha_i(\boldsymbol{x}_i)^d)\|^2\geq 0 \tag{1.2}$$

因此，原始 *Gram* 矩阵可以视做多个多项式核的和，当 $c\geq 0$ 时函数 $\kappa(\boldsymbol{x},\boldsymbol{y})$ 为核函数。
当 $c<0$ 时，若此时 d 为奇数，总能找到一对 $(\boldsymbol{x},\boldsymbol{y})$ 使得 $(\boldsymbol{x}^\top\boldsymbol{y}+c)^d$ 为负，此时该函数不满足 *Mercer* 条件，不为核函数。
(2). 当 $d=2$ 时，该核函数为 $\kappa(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}^\top\boldsymbol{y}+c)^2$。假设有 N 维空间的两组数据 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ 和 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_N\}$, 则*:*

$$\kappa(\boldsymbol{x},\boldsymbol{y})=(c+\boldsymbol{x}_1\boldsymbol{y}_1+\boldsymbol{x}_2\boldsymbol{y}_2+\cdots+\boldsymbol{x}_N\boldsymbol{y}_N)^2 \tag{1.3}$$

$$=c^2+\boldsymbol{x}_1^2\boldsymbol{y}_1^2+\boldsymbol{x}_2^2\boldsymbol{y}_2^2+\cdots+\boldsymbol{x}_N^2\boldsymbol{y}_N^2+2c(\boldsymbol{x}_1\boldsymbol{y}_1+\boldsymbol{x}_2\boldsymbol{y}_2+\cdots+\boldsymbol{x}_N\boldsymbol{y}_N) \tag{1.4}$$

$$+2c(\boldsymbol{x}_1\boldsymbol{y}_1\boldsymbol{x}_2\boldsymbol{y}_2+\boldsymbol{x}_1\boldsymbol{y}_1\boldsymbol{x}_3\boldsymbol{y}_3+\cdots+\boldsymbol{x}_1\boldsymbol{y}_1\boldsymbol{x}_N\boldsymbol{y}_N+\cdots) \tag{1.5}$$

$$=[c,\sqrt{2c}\boldsymbol{x}_1,\cdots,\sqrt{2c}\boldsymbol{x}_N,\boldsymbol{x}_1^2,\cdots,\boldsymbol{x}_N^2,\sqrt{2}\boldsymbol{x}_1\boldsymbol{x}_2,\cdots,\sqrt{2}\boldsymbol{x}_{N-1}\boldsymbol{x}_N,\cdots]^\top \tag{1.6}$$

$$[c,\sqrt{2c}\boldsymbol{y}_1,\cdots,\sqrt{2c}\boldsymbol{y}_N,\boldsymbol{y}_1^2,\cdots,\boldsymbol{y}_N^2,\sqrt{2}\boldsymbol{y}_1\boldsymbol{y}_2,\cdots,\sqrt{2}\boldsymbol{y}_{N-1}\boldsymbol{y}_N,\cdots] \tag{1.7}$$

$$=\phi(\boldsymbol{x})^\top\phi(\boldsymbol{y}) \tag{1.8}$$

因此，对于二阶多项式我们可以显式地得到它从输入空间到输出空间的映射。该 N 维数据被映射到一个 $(N+1)(N+2)/2$ 维的特征空间中。
若对 d 不加限制时, κ 将 N 维数据映射到了一个 C_{N+d}^d 维的特征空间中。*(*从上述 $d=2$ 的证明可以看出，映射后的特征空间的维数为一个多项式的和。*)*

2 [25pts] Dual Problem

给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, 其中 $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$. 对率回归 (logistic regression) 的优化问题为:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{m} \log\left(1 + e^{-y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2,$$

其中 $\lambda > 0$ 是超参数. 试推导上述问题的对偶问题.

Solution.

上述问题可改写为如下约束优化形式:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{m} \log\left(1 + e^{\xi_i}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$
$$s.t. \quad \xi_i = -y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}, \ for\ i \in [m]$$

引入 *Lagrange Multiplier* $\alpha = \{\alpha_i\}_{i\in[m]}$ 构造 *Lagrange Function:*

$$L(\xi, \boldsymbol{w}; \alpha) = \sum_{i=1}^{m} \log\left(1 + e^{\xi_i}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{m} \alpha_i(-\xi_i - y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w})$$

求偏导可得:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = \lambda \boldsymbol{w} - \displaystyle\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} & = 0 \\ \dfrac{\partial L}{\partial \xi_i} = \dfrac{e^{\xi_i}}{1 + e^{\xi_i}} - \alpha_i & = 0 \end{cases}$$

如上所示: 当 *Lagrange Function* 取极值点时: $\xi_i = \log(\frac{\alpha_i}{1-\alpha_i})$, $\boldsymbol{w} = \frac{1}{\lambda}\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}$

将上值代入 *Lagrange Function* 函数可得如下对偶问题:

$$\min_{\alpha} \frac{1}{2\lambda} \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \sum_{i=1}^{m} [\alpha_i \log \alpha_i + (1 - \alpha_i) \log(1 - \alpha_i)]$$
$$s.t. \quad 0 \le \alpha_i < 1, \ for\ i \in [m]$$

3 [25pts] Not PAC Learnable

(1) **[10 pts]** 是否存在可分但不 PAC 可学的假设空间? 证明你的结论.

(2) **[15 pts]** 是否存在可分、有限但不 PAC 可学的假设空间? 证明你的结论.

Proof.

(1). 存在可分但不 PAC 可学的假设空间。证明如下所示。

考虑一个简单的二分类问题，数据点是实数轴上的点，目标是根据点的位置分类。假设我们的假设空间只包含一个假设 $h(\boldsymbol{x}) = 1$，即对于任意的数据该假设总是预测正类。这个假设空间是可分的，因为我们可以构造一个训练集，使得所有点都是正类，这样 h 就可以完美地分类这个训练集。然而，这个假设空间不是 *PAC* 可学的，因为对于任何未见过的数据点 $\boldsymbol{x}$，h 都会预测正类，即使 $\boldsymbol{x}$ 实际上是负类。因此，h 在未见过的数据上表现不佳，无法以高概率正确分类大多数未见过的数据点。因此，这个假设空间是可分的，但不是 *PAC* 可学的。

(2). 有限可分假设空间一定是 PAC 可学的。证明如下所示。

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

$$P(h(\boldsymbol{x}) = \boldsymbol{y}) = 1 - P(h(\boldsymbol{x}) \neq \boldsymbol{y}) \tag{3.1}$$

$$= 1 - E(h) \tag{3.2}$$

$$< 1 - \epsilon \tag{3.3}$$

D 中的 m 个样本是从分布 $\mathcal{D}$ 中随机采样得到，因此，h 与 D 表现一致的概率为:

$$P((h(\boldsymbol{x}_1) = \boldsymbol{y}_1) \wedge \cdots \wedge (h(\boldsymbol{x}_m) = \boldsymbol{y}_m)) = (1 - P(h(\boldsymbol{x}_m) \neq \boldsymbol{y}_m))^m \tag{3.4}$$

$$< (1 - \epsilon)^m \tag{3.5}$$

由于学习算法输出的假设未知，但仅需保证泛化误差大于 ϵ 且在训练集上所有可分假设出现的概率之和不大于 δ 即可:

$$P(h \in \mathcal{H} : E(h) > \epsilon \wedge \hat{E}(h) = 0) < |\mathcal{H}|(1 - \epsilon)^m \tag{3.6}$$

$$< |\mathcal{H}|e^{-m\epsilon} \tag{3.7}$$

令 $|\mathcal{H}|e^{-m\epsilon} \leq \delta$ 即可保证 *PAC* 可学，即:

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

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. 接下来, 我们将证明一个更广泛的结论: 与坐标轴平行的超矩形是可学习的. 具体而言, 在 $\mathbb{R}^n$ 空间中, 一个与坐标轴平行的超矩形可以表示为 $[a_1, b_1] \times \cdots \times [a_n, b_n]$, 其中每一对 $[a_i, b_i]$ 定义了超矩形在第 i 维上的边界.

Proof.

如上规则所示, 在 $\mathbb{R}$ 空间中, 一个与坐标轴平行的超矩形可表示为一个线段, 不包括线段内部则可以将实数空间依照线段端点划分为两部分; 在 $\mathbb{R}^2$ 空间中, 一个与坐标轴平行的超矩形可表示为一个矩形, 不包括矩形内部则可以将实数空间依照矩阵的边划分为四部分; 在 $\mathbb{R}^3$ 空间中, 一个与坐标轴平行的超矩形可表示为一个平行六面体, 不包括平行六面体内部则可以将实数空间依照平行六面体的面划分为六部分; 同理, 在 $\mathbb{R}^n$ 空间中, 一个与坐标轴平行的超矩形可表示为一个 n 维超立方体, 不包括超立方体体内部则可以将实数空间依照超立方体的面划分为 $(2n)$ 部分。

依据书中对矩形 *PAC* 可学性的证明, 在此处做推广: 对于训练集 D, 学习算法 $\mathcal{L}$ 输出一个包含了 D 中所有正例的最小超立方体 R^D。令 $P(R)$ 表示目标概念的概率质量, 即分布内的点落到目标概念的概率。由于学习算法 $\mathcal{L}$ 的错误仅可能出现在 R 内的点上, 可以设 $P(R) > \epsilon$。因为该超立方体可以划分为 $(2n)$ 部分, 若其泛化误差 $E(R^D) > \epsilon$, 则 R^D 必然至少与超立方体外的某一部分不相交, 此时训练集 D 的样本出现在超立方体外的某一部分的概率为 $\epsilon/2n$。设 D 包含 m 个样本, 则:

$$P_{D\in\mathcal{D}^m}(E(R^D) > \epsilon) \le P_{D\in\mathcal{D}^m}(\cup_{i=1}^{2n}\{R^D \cap r_i = \emptyset\}) \tag{4.1}$$

$$\le \sum_{i=1}^{2n} P_{D\in\mathcal{D}^m}(\{R^D \cap r_i = \emptyset\}) \tag{4.2}$$

$$\le 2n(1-\epsilon/2n)^m \tag{4.3}$$

$$\le 2ne^{-\epsilon m/2n} \tag{4.4}$$

令 $2ne^{-\epsilon m/2n} \le \delta$ 即可确保:

$$P_{D\in\mathcal{D}^m}(E(R^D) \le \epsilon) = 1 - P_{D\in\mathcal{D}^m}(E(R^D) > \epsilon) \ge 1 - \delta$$

此时:

$$m \ge \frac{2n}{\epsilon}\log\frac{2n}{\delta}$$