机器学习导论
习题四参考答案

学号, 作者姓名, 邮箱

2020 年 5 月 16 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是署你名字的工作必须由你完成，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。对于完成作业中有关键作用的公开资料，应予以明显引用；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，抄袭和被抄袭双方的成绩都将被取消。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在 LaTeX 模板中第一页填写个人的姓名、学号、邮箱信息；

(2) 本次作业需提交该 pdf 文件、问题 4 可直接运行的源码 (main.py)、问题 4 的输出文件 (学号 _ypred.csv)，将以上三个文件压缩成 zip 文件后上传。zip 文件格式为**学号.zip**，例如 170000001.zip；pdf 文件格式为**学号 _ 姓名.pdf**，例如 170000001_张三.pdf。

(3) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数；

(4) 本次作业提交截止时间为**5 月 7 日 23:59:59**。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

[30 pts] Problem1 1 [Kernel Functions]

(1) **[10 pts]** 对于 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$，考虑函数 $\kappa(x, y) = \tanh(a\boldsymbol{x}^\top \boldsymbol{y} + b)$，其中 a, b 是任意实数。试说明 $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件。

(2) **[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。

(3) **[10 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) 考虑 $m = 1$ 的情况，令 $\boldsymbol{x} = (x, 0, \cdots, 0)$，则核矩阵为 $\boldsymbol{K} = [\tanh(ax^2 + b)]$。由于 $\tanh y \geq 0$ 当且仅当 $y \geq 0$，可知 κ 为核函数的必要条件是 $ax^2 + b \geq 0$。下面分情况举反例。

- 当 $a < 0$ 且 $b \leq 0$ 时，取 $x = 1$，则有 $ax^2 + b = a + b < 0$。
- 当 $a < 0$ 且 $b > 0$ 时，取 $x = \sqrt{-2b/a}$，则有 $ax^2 + b = -b < 0$。
- 当 $a \geq 0$ 且 $b < 0$ 时，取 $x = 0$，则有 $ax^2 + b = b < 0$。

(2) 考虑课本中定理 6.1。

- 当 $c \geq 0$ 时，下证 κ 是核函数。由于 κ 是 d 个 $\boldsymbol{x}^\top \boldsymbol{y} + c$ 的直积，由课本式 (6.26)，只需证明 $\boldsymbol{x}^\top \boldsymbol{y} + c$ 是核函数即可。由于 $\boldsymbol{x}^\top \boldsymbol{y}$ 与 c 显然是核函数，由课本式 (6.25)，知 $\boldsymbol{x}^\top \boldsymbol{y} + c$ 确为核函数。
- 当 $c < 0$ 时，取 $m = 2$，$\boldsymbol{x} = (\sqrt{-2c}, 0, \cdots, 0)$ 与 $\boldsymbol{y} = (-\sqrt{-2c}, 0, \cdots, 0)$，则核矩阵为

$$\boldsymbol{K} = \begin{bmatrix} (-c)^d & (3c)^d \\ (3c)^d & (-c)^d \end{bmatrix} \tag{1}$$

  其行列式为 $|\boldsymbol{K}| = (1 - 3^{2d})c^{2d} < 0$，即核矩阵非半正定，故 κ 不是核函数。

综上，当 $c \geq 0$ 时，κ 是核函数；当 $c < 0$ 时，κ 不是核函数。

(3) $d = 2$ 时，κ 将 N 维数据映射到 $\binom{N+2}{2}$ 维空间中。令 $\boldsymbol{x} = (x_1, \cdots, x_N)$，则映射函数为

$$\phi(\boldsymbol{x}) \in \left(\times_{i=1}^{N} \{x_i^2\}\right) \times \left(\times_{i=1}^{N} \times_{j=i+1}^{N} \{\sqrt{2}x_i x_j\}\right) \times \left(\times_{i=1}^{N} \{\sqrt{2c}x_i\}\right) \times \{c\}, \tag{2}$$

其中 $\times$ 为笛卡尔积，由于此笛卡尔积得到的集合中仅一个元素，ϕ 等于该元素。更一般的，κ 将 N 维数据映射到 $\binom{N+d}{d}$ 维空间中。

助教反馈

- 第一题作为原创题均分较低，主要问题在于判定核函数过程中的逻辑使用，得分较低的同学可以仔细阅读答案和反馈。
- P1.1：半正定矩阵主子式非负，只能得到 $\kappa(x, x) \geq 0$，无法得到 $\kappa(x, y) \geq 0$，后者事实上对于半正定矩阵不成立。

- P1.1：要证的是必要条件，有部分同学得到核函数的一个必要条件（记为 C）后，说 $a \geq 0, b \geq 0$ 可以推到出 C。这一论证过程只说明了 $a \geq 0, b \geq 0$ 是一个必要条件的充分条件，不得分。
- P1.2：c, d, N 均需讨论，下述论证不成立：d 为奇数时，$c < 0$ 时不是核函数，从而 $c < 0$ 是不是核函数。
- P1.2：有不少同学构造了一个映射函数，然后根据 $c < 0$ 时映射函数中出现虚数来说明不是核函数，这一论证不成立。核函数只需要存在一个映射即可，给出的映射不成立，可能存在其他映射。
- P1.2：证明一个函数不是核函数，可以随意设置样本数举反例；但说明一个函数是核函数时，应论证任意样本数下均成立。
- P1.3：(共性问题)映射函数的书写鲜有严谨而直观的作业。严谨：使用省略号来表示较为复杂的结构是不够严谨的，可能会产生歧义，比如把混合项写为 $(x_1x_2, \cdots, x_{N-1}x_N)$，可以误解为 $(x_ix_{i+1})_{i=1}^{N-1}$。直观：Intro to ML 中的写法严谨，但对于 $d = 2$ 可以更加直观。这里只对上述举例的写法扣分，其余写法均不扣分。
- P1.3：(较为共性的问题)很多同学提到了无序单项式空间这一概念，有两点需要注意。引用：据了解，17 级 cs 与 18 级 ai 都没有学过这一概念，网上严谨的定义也很少，使用这些概念最好添加引用。正确性：这里只是普通的 $\mathbb{R}^n$ 空间，与单项式空间没有关系；并且映射到单项式空间的含义是 $\phi(x)$ 的结果为单项式空间中的一个单项式，而这里是一个向量。
- P1.3：注意符号使用，数据维数是 N，很多同学用了 n 或者 k，都酌情扣分。
- P1.3：很多同学讨论了 $c = 0$ 和 $c \neq 0$，这一讨论很好，但不做要求。

[30 pts] Problem 2 [Surrogate Function in SVM]

在软间隔支持向量机问题中，我们的优化目标为

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) - 1\right). \tag{3}$$

然而 $\ell_{0/1}$ 数学性质不太好，它非凸、非连续，使得式 (3) 难以求解。实践中我们通常会将其替换为“替代损失”，替代损失一般是连续的凸函数，且为 $\ell_{0/1}$ 的上界，比如 hinge 损失，指数损失，对率损失。下面我们证明在一定的条件下，这样的替换可以保证最优解不变。

我们考虑实值函数 $h : \mathcal{X} \to \mathbb{R}$ 构成的假设空间，其对应的二分类器 $f_h : \mathcal{X} \to \{+1, -1\}$ 为

$$f_h(x) = \begin{cases} +1 & \text{if } h(x) \geq 0 \\ -1 & \text{if } h(x) < 0 \end{cases}$$

h 的期望损失为 $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[I_{f_h(x)\neq y}\right]$，其中 I 为指示函数。设 $\eta(x) = \mathbb{P}(y = +1|x)$，则贝叶斯最优分类器当 $\eta(x) \geq \frac{1}{2}$ 时输出 1，否则输出 -1。因此可以定义贝叶斯得分 $h^*(x) = \eta(x) - \frac{1}{2}$ 和贝叶斯误差 $R^* = R(h^*)$。

设 $\Phi:\mathbb{R}\to\mathbb{R}$ 为非减的凸函数且满足 $\forall u\in\mathbb{R}, 1_{u\le 0}\le\Phi(-u)$。对于样本 (x,y)，定义函数 h 在该样本的 Φ-损失为 $\Phi(-yh(x))$，则 h 的期望损失为 $\mathcal{L}_\Phi(h)=\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\Phi(-yh(x))]$。定义 $L_\Phi(x,u)=\eta(x)\Phi(-u)+(1-\eta(x))\Phi(u)$，设 $h_\Phi^*(x)=\mathop{\operatorname{argmin}}_{u\in[-\infty,+\infty]} L_\Phi(x,u)$，$\mathcal{L}_\Phi^*=\mathcal{L}_\Phi(h_\Phi^*(x))$。

我们考虑如下定理的证明：

若对于 Φ，存在 $s\ge 1$ 和 $c>0$ 满足对 $\forall x\in\mathcal{X}$ 有

$$|h^*(x)|^s=\left|\eta(x)-\frac{1}{2}\right|^s\le c^s\left[L_\Phi(x,0)-L_\Phi\left(x,h_\Phi^*(x)\right)\right] \tag{4}$$

则对于任何假设 h，有如下不等式成立

$$R(h)-R^*\le 2c\left[\mathcal{L}_\Phi(h)-\mathcal{L}_\Phi^*\right]^{\frac{1}{s}} \tag{5}$$

(1) **[5 pts]** 请证明

$$\Phi\left(-2h^*(x)h(x)\right)\le L_\Phi(x,h(x)) \tag{6}$$

(2) **[10 pts]** 请证明

$$R(h)-R^*=2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[|h^*(x)|\,1_{h(x)h^*(x)\le 0}\right] \tag{7}$$

提示：先证明

$$R(h)=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[2h^*(x)1_{h(x)<0}+(1-\eta(x))\right]$$

(3) **[10 pts]** 利用式 (6) 和式 (7) 完成定理的证明。

(4) **[5 pts]** 请验证对于 Hinge 损失 $\Phi(u)=\max(0,1+u)$，有 $s=1,c=\frac{1}{2}$。

Solution.

(1) 式 (6) 证明如下：

$$\begin{aligned}\Phi\left(-2h^*(x)h(x)\right)&=\Phi((1-2\eta(x))h(x))\\&=\Phi(\eta(x)(-h(x))+(1-\eta(x))h(x))\\&\le\eta(x)\Phi((-h(x)))+(1-\eta(x))\Phi(h(x))=L_\Phi(x,h(x))\end{aligned}$$

(2) 式 (7) 证明如下：

首先，我们证明 $R(h)=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[2h^*(x)1_{h(x)<0}+(1-\eta(x))\right]$

$$\begin{aligned}R(h)&=\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[I_{f_h(x)\ne y}\right]\\&=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\eta(x)I_{h(x)<0}+(1-\eta(x))I_{h(x)>0}+(1-\eta(x))I_{h(x)=0}\right]\\&=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\eta(x)I_{h(x)<0}+(1-\eta(x))I_{h(x)\ge 0}\right]\\&=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\eta(x)I_{h(x)<0}+(1-\eta(x))\left(1-I_{h(x)<0}\right)\right]\\&=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[[2\eta(x)-1]I_{h(x)<0}+(1-\eta(x))\right]\\&=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[2h^*(x)I_{h(x)<0}+(1-\eta(x))\right]\end{aligned}$$

于是有

$$\begin{aligned} R(h) - R\left(h^*\right) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[2\left[h^*(x)\right]\left(I_{h(x)\le 0} - I_{h^*(x)\le 0}\right)\right] \\ &= 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\left|h^*(x)\right| I_{h(x)h^*(x)\le 0}\right] \end{aligned}$$

最后一个等式可以通过对 $h(x), h^*(x)$ 的符号进行讨论验证。

(3) 定理证明如下：

$$\begin{aligned} & R(h) - R\left(h^*\right) \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[|2\eta(x)-1| I_{h(x)h^*(x)\le 0}\right] \text{ (式 (7))} \\ &\le \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}|2\eta(x)-1|^s I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \text{ (琴生不等式)} \\ &\le 2c\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\Phi(0) - L_\Phi\left(x, h_\Phi^*(x)\right)\right] I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \text{ (假设)} \\ &\le 2c\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[\Phi(-2h^*(x)h(x)) - L_\Phi\left(x, h_\Phi^*(x)\right)\right] I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \text{ (Φ 非减)} \\ &\le 2c\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[L_\Phi\left(x, h(x)\right) - L_\Phi\left(x, h_\Phi^*(x)\right)\right] I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \text{ (式 (6))} \\ &\le 2c\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[L_\Phi\left(x, h(x)\right) - L_\Phi\left(x, h_\Phi^*(x)\right)\right]\right]^{\frac{1}{s}} \\ &= 2c\left[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*\right]^{\frac{1}{s}} \end{aligned}$$

(4) 当 $\Phi(u) = \max(0, 1+u)$ 时，

$$\begin{aligned} L_\Phi(x,u) =& \eta(x)\max(0,1-u) + (1-\eta(x))\max(0,1+u) \\ =& \begin{cases} (1-u)\eta(x) & u < -1 \\ 1 + (1-2\eta(x))u & -1 \le u < 1 \\ (1+u)(1-\eta(x)) & u \ge 1 \end{cases} \end{aligned} \tag{8}$$

由于 $h_\Phi^*(x) = \mathop{\operatorname{argmin}}_{u\in[-\infty,+\infty]} L_\Phi(x,u)$，可得当 $\eta(x) > \frac{1}{2}$ 时，$h_\Phi^*(x) = 1, L_\Phi\left(x, h_\Phi^*(x)\right) = 2(1-\eta(x))$；当 $\eta(x) \le \frac{1}{2}$ 时，$h_\Phi^*(x) = -1, L_\Phi\left(x, h_\Phi^*(x)\right) = 2\eta(x)$。又 $L_\Phi(x,0) = \Phi(0) = 1$，可验证式 (4) 成立，$s = 1, c = \frac{1}{2}$，且不等式取等。

助教反馈

- 这道题总体而言比较烦琐，主要是想让同学体会一下如下思想：对于使用了 $0-1$loss 的 ERM 或者 SRM 问题直接进行优化是 NP-难的问题，一种常见的技巧就是寻找一些满足特定性质的替代损失函数代替原有的 $0-1$ 损失函数，而这样一种技巧可以保证最优解不变，即使用替代损失函数得到的最优解同时也是原始问题的最优解。值得注意的是，定理并不仅仅针对于 SVM 问题，所有采用了替代损失函数的 ERM 或者 SRM 的问题都有一样的定理结果，SVM 只是其中一个具体的实例。
- 在证明过程中主要用到了凸函数的性质，通过分类讨论合并示性函数，琴生不等式。在验证过程中，求解出分段线性函数 $L_\Phi(x,u)$ 及相应的极值即可完成验证。
- 大家的答题情况很好，批完作业的助教献上一张表情包

[20 pts] Problem 3 [Generalization Error of SVM]

留一损失 (leave-one-out error) 使用留一法对分类器泛化错误率进行估计，即：每次使用一个样本作为测试集，剩余样本作为训练集，最后对所有测试误差求平均。对于 SVM 算法 $\mathcal{A}$，令 h_S 为该算法在训练集 S 上的输出，则该算法的经验留一损失可形式化定义为

$$\hat{R}_{\mathrm{LOO}}(\mathcal{A}) = \frac{1}{m}\sum_{i=1}^{m} 1_{h_{S-\{x_i\}}(x_i)\neq y_i}. \tag{9}$$

本题通过探索留一损失的一些数学性质，来分析 SVM 的泛化误差，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。)

(1) **[10pts]** 在实践中，测试误差相比于泛化误差是很容易获取的。虽然测试误差不一定是泛化误差的准确估计，但测试误差与泛化误差往往能在期望意义下一致。试证明留一损失满足该性质，即

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] = \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[R(h_{S'})]. \tag{10}$$

(2) **[5 pts]** SVM 之所以取名为 SVM，是因为其训练结果仅与一部分样本 (即支持向量) 有关。这一现象可以抽象的表示为，如果 x 不是 h_S 的支持向量，则 $h_{S-\{x\}} = h_S$。这一性质在分析误差时有关键作用，考虑如下问题：如果 x 不是 h_S 的支持向量，$h_{S-\{x\}}$ 会将 x 正确分类吗，为什么？该问题结论的逆否命题是什么？

(3) **[5 pts]** 基于上一小问的结果，试证明下述 SVM 的泛化误差界

$$\mathbb{E}_{S\sim\mathcal{D}^m}[R(h_S)] \leq \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right], \tag{11}$$

其中 $N_{SV}(S)$ 为 h_S 支持向量的个数。

Solution.

(1) 证明如下：

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] =& \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}[1_{h_{S-\{x_i\}}(x_i)\neq y_i}] \\
=& \mathbb{E}_{S\sim\mathcal{D}^m}[1_{h_{S-\{x_1\}}(x_1)\neq y_1}] \\
=& \mathbb{E}_{S'\sim\mathcal{D}^{m-1},x_1\sim\mathcal{D}}[1_{h_{S'}(x_1)\neq y_1}] \\
=& \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}\left[\mathbb{E}_{x_1\sim\mathcal{D}}[1_{h_{S'}(x_1)\neq y_1}]\right] \\
=& \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[R(h_{S'})].
\end{aligned} \tag{12}$$

(2) 因为是可分情形，所以 h_S 可以将 x 分类正确，又 $h_{S-\{x\}} = h_S$，故 $h_{S-\{x\}}$ 也可以将 x 分类正确。逆否命题为：如果 $h_{S-\{x\}}$ 将 x 分类错误，则 x 是 h_S 的支持向量。

(3) 令 S 为含有 $m+1$ 个样本的样本集，则在 S 上使用留一法时，每一个错分类样本都是 h_S 的支持向量。故 S 上的留一损失小于等于 $N_{SV}(S)/(m+1)$，从而

$$\mathbb{E}_{S\sim\mathcal{D}^m}[R(h_S)] = \mathbb{E}_{S\sim\mathcal{D}^{m+1}}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] \leq \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right]. \tag{13}$$

助教反馈

- P3.1：变量较为复杂时，期望的下标不宜省略。
- P3.1：注意符号的意义，每个符号都需有固定的含义。其一，在将 $1/m\cdot\sum_{i=1}^{m}$ 变为 1 后，期望表达式中的 i 变为了无意义的符号，需要特殊处理或说明。其二，新引入的符号需要说明含义。
- tex 的注意事项：不少同学将 $S-\{x\}$ 写成了 $S-x$，这一写法是不严谨的，tex 中 {} 作为特殊符号需注意。

[20 pts] Problem 4 [NN in Practice]

请结合编程题指南进行理解

在训练神经网络之前，我们需要确定的是整个网络的结构，在确定结构后便可以输入数据进行端到端的学习过程。考虑一个简单的神经网络：输入是 2 维向量，隐藏层由 2 个隐层单元组成，输出层为 1 个输出单元，其中隐层单元和输出层单元的激活函数都是 *Sigmoid* 函数。请打开 **main.py** 程序并完成以下任务：

(1) **[4 pts]** 请完成 Sigmoid 函数及其梯度函数的编写。

(2) **[2 pts]** 请完成 MSE 损失函数的编写。

(3) **[9 pts]** 请完成 NeuralNetwork_221() 类中 train 函数的编写，其中包括向前传播 (可参考 predict 函数)、梯度计算、更新参数三个部分。

(4) **[5 pts]** 请对测试集 (test_feature.csv) 所提供的数据特征完成尽量准确的分类预测。

Solution. 此处用于写解答 (中英文均可)