机器学习导论
习题四

171830635, 俞星凯, yuxk@smail.nju.edu.cn

2020 年 5 月 13 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 14 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] Problem 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) 因为对于任意数据集，核矩阵都是半正定的，所以对于任意 $x \in \mathbb{R}^N$，满足 $\kappa(x, x) \geq 0$，即 $ax^T x + b \geq 0$，而 $x^T x \geq 0$，所以 $a \geq 0, b \geq 0$，于是 $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件。

(2) 当 $c \geq 0$ 时，使用二项式定理可以将 $\kappa(x, y) = (x^T y + c)^d$ 展开为 $n + 1$ 个项的线性组合，并且系数非负。已知 $\kappa'(x, y) = (x^T y)^k$ 对任意正整数 k 都是核函数，根据核函数的非负线性组合也是核函数，κ 是核函数。
当 $c < 0$ 时，考虑 d 的奇偶性。如果 d 是奇数，根据 $\kappa(\mathbf{0}_N, \mathbf{0}_N) = c^d < 0$，$\kappa$ 不是核函数。如果 d 是偶数，考虑数据集 $D = \{x, y\}$，核矩阵 K 的行列式 $det(K) = (x^T x + c)^d (y^T y + c)^d - (x^T y + c)^{2d}$，令 $x = \mathbf{0}_N, y = (\mathbf{0}_{N-1}; \sqrt{-c})$，则 $x^T x = 0, x^T y = 0, y^T y = -c$，于是 $det(K) = -c^{2d} < 0$，κ 不是核函数。
综上所述，当 $c \geq 0$ 时，κ 是核函数；当 $c < 0$ 时，κ 不是核函数。

(3) κ 将 N 维数据映射到了 $\frac{(N+1)(N+2)}{2}$ 维空间，映射函数

$$\kappa(x) = (c, \sqrt{2c}x_1, \sqrt{2c}x_2, \cdots, \sqrt{2c}x_N, x_1^2, x_2^2, \cdots, x_N^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3 \cdots, \sqrt{2}x_{N-1}x_N)$$

考虑 $\kappa(x, y) = (x^T y + c)^d$ 的展开式的任意一项中，$c, x_1y_1, x_2y_2, \cdots, x_Ny_N$ 的指数分别是 $d_0, d_1, d_2, \cdots, d_N$，满足 $\sum_{i=0}^N d_i = d$，并且 $d_i (0 \leq i \leq N)$ 是非负整数。由排列组合知识易知有 C_{N+d}^d 种可能，即 κ 将 N 维数据映射到了 C_{N+d}^d 维空间。

[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

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

然而 $\ell_{0/1}$ 数学性质不太好，它非凸、非连续，使得式 (1) 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 $\forall u \in \mathbb{R}, 1_{u\leq 0} \leq \Phi(-u)$。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 $s \geq 1$ 和 $c > 0$ 满足对 $\forall x \in \mathcal{X}$ 有

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

则对于任何假设 h，有如下不等式成立

$$R(h) - R^* \leq 2c\left[\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi\right]^{\frac{1}{s}} \tag{3}$$

(1) **[5 pts]** 请证明

$$\Phi\left(-2h^*(x)h(x)\right) \leq L_\Phi(x, h(x)) \tag{4}$$

(2) **[10 pts]** 请证明

$$R(h) - R^* \leq 2 \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[|h^*(x)|\, 1_{h(x)h^*(x)\leq 0}\right] \tag{5}$$

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 (4) 和式 (5) 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) 将 $h^*(x) = \eta(x) - \frac{1}{2}$ 代入 $\Phi(-2h^*(x)h(x))$ 得

$$\Phi(-2h^*(x)h(x)) = \Phi((1 - 2\eta(x))h(x)) = \Phi(-\eta(x)h(x) + (1 - \eta(x))h(x))$$

因为 Φ 是凸函数，由凸函数的性质得

$$\Phi(\eta(x)(-h(x)) + (1 - \eta(x))h(x)) \leq \eta(x)\Phi(-h(x)) + (1 - \eta(x))\Phi(h(x)) = L_\Phi(x, h(x))$$

(2) 计算 $R(h)$

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

将 h^* 代入

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

所以

$$\begin{aligned}
R(h) - R^* &= 2\ \mathbb{E}_{x\sim\mathcal{D}_x}[h^*(x)(1_{h(x)<0} - 1_{h^*(x)<0})] \\
&\leq 2\ \mathbb{E}_{x\sim\mathcal{D}_x}[|h^*(x)||1_{h(x)<0} - 1_{h^*(x)<0}|] \\
&= 2\ \mathbb{E}_{x\sim\mathcal{D}_x}[|h^*(x)|1_{h(x)h^*(x)\leq 0}]
\end{aligned}$$

(3) 计算 $\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*$

$$\begin{aligned}
\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^* &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\Phi(-yh(x))] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[\Phi(-yh_\Phi^*(x))] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}[\Phi(-h(x))\eta(x) + \Phi(h(x))(1-\eta(x))] \\
&\quad + \mathbb{E}_{x\sim\mathcal{D}_x}[\Phi(-h_\Phi^*(x))\eta(x) + \Phi(h_\Phi^*(x))(1-\eta(x))] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}[L_\Phi(x,h(x)) - L_\Phi(x,h_\Phi^*(x))]
\end{aligned}$$

由式 (2) 和式 (5) 得

$$\begin{aligned}
R(h) - R^* &\leq 2\ \mathbb{E}_{x\sim\mathcal{D}_x}[|h^*(x)|1_{h(x)h^*(x)\leq 0} \\
&\leq 2\ \mathbb{E}_{x\sim\mathcal{D}_x}[c[L_\Phi(x,0) - L_\Phi(x,h_\Phi^*(x))]^{\frac{1}{s}}1_{h(x)h^*(x)\leq 0}] \\
&= 2c\ \mathbb{E}_{x\sim\mathcal{D}_x}[[(L_\Phi(x,0) - L_\Phi(x,h_\Phi^*(x)))1_{h(x)h^*(x)\leq 0}]^{\frac{1}{s}}]
\end{aligned}$$

要证式 (3)，只要证明

$$(L_\Phi(x,0) - L_\Phi(x,h_\Phi^*(x)))1_{h(x)h^*(x)\leq 0} \leq L_\Phi(x,h(x)) - L_\Phi(x,h_\Phi^*(x))$$

分类讨论，若 $h(x)h^*(x) > 0$

$$L_\Phi(x,h(x)) - L_\Phi(x,h_\Phi^*(x)) \geq L_\Phi(x,h(x)) - \min_u L_\Phi(x,u) \geq 0$$

若 $h(x)h^*(x) \leq 0$

$$\begin{aligned}
L_\Phi(x,h(x)) - L_\Phi(x,h_\Phi^*(x)) &\geq \Phi(-2h^*(x)h(x)) - L_\Phi(x,h_\Phi^*(x)) \\
&\geq \Phi(0) - L_\Phi(x,h_\Phi^*(x)) \\
&= L_\Phi(x,0) - L_\Phi(x,h_\Phi^*(x))
\end{aligned}$$

因此

$$R(h) - R^* \leq 2c[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*]^{\frac{1}{s}}$$

(4) 将 $L_\Phi(x,u)$ 和 $\Phi(u)$ 展开

$$L_\Phi(x,h_\Phi(x)) = \eta(x)\max(0,1-h_\Phi(x)) + (1-\eta(x))\max(0,1+h_\Phi(x))$$

因为 $L_\Phi(x,h^*_\Phi(x))$ 是最小的，下面分段求出 $L_\Phi(x,h^*_\Phi(x))$
当 $h_\Phi(x) \geq 1$ 时

$$L_\Phi(x,h^*_\Phi(x)) = (1-\eta(x))(1+h_\Phi(x)) \geq 2(1-\eta(x))$$

当 $h_\Phi(x) \leq -1$ 时

$$L_\Phi(x,h_\Phi(x)) = \eta(x)(1-h_\Phi(x)) \geq 2\eta(x)$$

当 $-1 \leq h_\Phi(x) \leq 1$ 时

$$L_\Phi(x,h_\Phi(x)) = \eta(x)(1-h_\Phi(x)) + (1-\eta(x))(1+h_\Phi(x)) = 1 + (1-2\eta(x))h_\Phi(x)$$

若 $1-2\eta(x) \geq 0$

$$L_\Phi(x,h_\Phi(x)) \geq 2\eta(x)$$

若 $1-2\eta(x) \leq 0$

$$L_\Phi(x,h_\Phi(x)) \geq 2(1-\eta(x))$$

因此得到

$$L_\Phi(x,h^*_\Phi(x)) = \min L_\Phi(x,h_\Phi(x)) = \min(2\eta(x), 2(1-\eta(x)))$$

所以

$$\begin{aligned}
c^s[L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x))] &= \frac{1}{2}[L_\Phi(x,0) - L_\Phi(x,h^*_\Phi(x))] \\
&= \frac{1}{2}[\Phi(0) - \min(2\eta(x), 2(1-\eta(x)))] \\
&= \frac{1}{2}\max(1-2\eta(x), -1+2\eta(x)) \\
&= \frac{1}{2}|(2\eta(x)-1)| \\
&= |\eta(x) - \frac{1}{2}| \\
&= |h^*(x)|^s
\end{aligned}$$

[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

$$\hat{R}_{\mathrm{LOO}}(\mathcal{A}) = \frac{1}{m}\sum_{i=1}^{m} 1_{h_{S-\{x_i\}}(x_i)\neq y_i}. \tag{6}$$

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

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] = \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[R(h_{S'})]. \tag{7}$$

(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

$$\mathbb{E}_{S\sim\mathcal{D}^m}[R(h_S)] \leq \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right], \tag{8}$$

其中 $N_{SV}(S)$ 为 h_S 支持向量的个数。

Solution.

(1)

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] &= \mathbb{E}_{S\sim\mathcal{D}^m}[\mathbb{E}_{x\in S}[1_{h_{S-\{x\}}(x)\neq y}]] \\
&= \mathbb{E}_{S\sim\mathcal{D}^m, x\in S}[1_{h_{S-\{x\}}(x)\neq y}] \\
&= \mathbb{E}_{S'\sim\mathcal{D}^{m-1}, x\sim D}[1_{h_{S'}(x)\neq y}] \\
&= \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[\mathbb{E}_{x\sim D}[1_{h_{S'}(x)\neq y}]] \\
&= \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[R(h_{S'})]
\end{aligned}$$

(2) 因为 $h_{S-\{x\}} = h_S$，并且仅考虑可分情形，所以 $h_{S-\{x\}}$ 会将 x 正确分类。逆反命题是如果 $h_{S-\{x\}}$ 将 x 错误分类，那么 x 是 h_S 的支持向量。

(3) 如果 x 不是 h_S 的支持向量，$h_{S-\{x\}}$ 会将 x 正确分类；如果 x 是 h_S 的支持向量，$h_{S-\{x\}}$ 对 x 分类可能正确也可能错误。所以分类错误个数少于支持向量个数。

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}[R(h_S)] &= \mathbb{E}_{S\sim\mathcal{D}^{m+1}}[\hat{R}_{\mathrm{LOO}}(\mathcal{A})] \\
&= \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{1}{m+1}\sum_{i=1}^{m+1} 1_{h_{S-\{x_i\}}(x_i)\neq y_i}\right] \\
&\leq \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right]
\end{aligned}$$

[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.

(1) Sigmoid 的函数 $f(x) = \frac{1}{1+e^{-x}}$，其导数 $f'(x) = f(x)(1 - f(x))$。

(2) MSE 损失函数 $L(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$。

(3) 前向传播过程是，$h_1 = sigmoid(w_1x_1 + w_2x_2 + b_1)$，$h_2 = sigmoid(w_3x_1 + w_4x_2 + b_2)$，$ol = sigmoid(w_5h_1 + w_6h_2 + b_3)$。
反向传播先计算 $\frac{\partial L}{\partial ol}$，这与损失函数 L 的形式有关，对于 MSE 为 $\hat{y} - y$。然后计算 $\frac{\partial ol}{\partial w_5} = \frac{\partial ol}{\partial sum\_ol}\frac{\partial sum\_ol}{\partial w_5} = ol(1 - ol)h_1$，同理 $\frac{\partial ol}{\partial w_6} = ol(1 - ol)h_2$，$\frac{\partial ol}{\partial b_3} = ol(1 - ol)$。接着计算 $\frac{\partial ol}{\partial h_1} = ol(1-ol)w_5$，$\frac{\partial ol}{\partial h_2} = ol(1-ol)w_6$。这时，再求 $\frac{\partial h_1}{\partial w_1}$ 等的过程就与求 $\frac{\partial ol}{\partial h_1}$ 类似，不再赘述。得到了每一层的输出对输入的偏导，只需使用链式法则相乘，即得损失对参数的偏导，再乘学习率，就是每轮更新的大小。

(4) 笔者调整了学习率，使用交叉验证确定早停轮数。同时还尝试了交叉熵损失，但效果并未显著提升，最终还是使用 MSE 损失。