机器学习导论
习题四

171860607, 白晋斌, 810594956@qq.com

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. 此处用于写解答 (中英文均可)

(1) 非线性支持向量机的对偶问题为

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0, i = 1, 2, \cdots, m \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, m \end{aligned} \tag{1}$$

其中, 核函数

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \tag{2}$$

$Mercer$ 定理说明了 κ 是核函数当且仅当核矩阵 K 总是半正定的. 即 $Mercer$ 定理要求

$$\int_C K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \tag{3}$$

κ 为半正定核需要满足

$$\begin{aligned} \kappa(||\mathbf{x}||^2) &\geq 0 \\ \kappa'(||\mathbf{x}||^2) &\geq 0 \\ \kappa'(||\mathbf{x}||^2) + ||\mathbf{x}||^2 \kappa''(||\mathbf{x}||^2) &\geq 0 \end{aligned} \tag{4}$$

其中第一个不等式源于 (3), 第二个和第三个不等式源于半正定矩阵的特征值非负的性质.
因此 $\kappa(x, y) = \tanh(a\boldsymbol{x}^\top \boldsymbol{y} + b)$ 为半正定核需要满足

$$\begin{aligned} b &\geq 0 \\ a &\geq 0 \\ 1 - 2a||\mathbf{x}||^2 \tanh(a||\mathbf{x}||^2 + b) &\geq 0 \end{aligned} \tag{5}$$

其中第一个不等式可以对 (4) 中第一个不等式 $||\boldsymbol{x}||^2$ 取极小值获得. 第二、三个不等式依次由 (4) 中第二、三个不等式获得.
当 κ 为核函数, 式 (2)、式 (4)、式 (5) 依次成立, 从而推出 $a \geq 0, b \geq 0$. 因此,$a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件.
反正, 当 $a \geq 0, b \geq 0$ 时, 式 (5) 未必成立, κ 未必为核函数, 因此 $a \geq 0, b \geq 0$ 不是 κ 为核函数的充分条件.
因此, $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件.

(2) 从第一问可以得知, $Mercer$ 定理说明了 κ 是核函数当且仅当核矩阵 K 总是半正定的. 即 $Mercer$ 定理要求

$$\int_C K(\mathbf{x},\mathbf{y})g(\mathbf{x})g(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0 \tag{6}$$

κ 为半正定核需要满足

$$\begin{aligned}\kappa(||\mathbf{x}||^2) &\geq 0 \\ \kappa'(||\mathbf{x}||^2) &\geq 0 \\ \kappa'(||\mathbf{x}||^2) + ||\mathbf{x}||^2\kappa''(||\mathbf{x}||^2) &\geq 0\end{aligned} \tag{7}$$

其中第一个不等式源于 (6), 第二个和第三个不等式源于半正定矩阵的特征值非负的性质.
此外,$\kappa(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y}+c)^d$ 可以看作是 d 个 $\kappa^{'}(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}+c$ 直积所得.
因此 $\kappa(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y}+c)^d$ 为半正定核只需要满足 $\kappa^{'}(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}+c$ 为半正定核.
将 $\kappa^{'}(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}+c$ 代入 (7) 得 $c \geq 0$, 该不等式由 (7) 中对第一个不等书 $||\mathbf{x}^2||$ 取极小值获得, (7) 中第二、三个不等式代入后恒成立.
综上所述, 当 $c \geq 0$ 时, 函数 κ 为核函数, 当 $c < 0$ 时, 函数 κ 不是核函数.

(3) 考虑 $d = 2$ 的情况对 $\kappa(\boldsymbol{x},\boldsymbol{y})$ 进行展开:

$$\begin{aligned}\kappa(\boldsymbol{x},\boldsymbol{y}) &= (\boldsymbol{x}^T\boldsymbol{y}+c)^2 \\ &= (\sum_{i=1}^{N} x_i y_i + c)^2 \\ &= \sum_{i=1}^{N}(x_i^2)(y_i^2) + \sum_{i=2}^{N}\sum_{j=1}^{i-1}(\sqrt{2}x_ix_j)(\sqrt{2}y_iy_j) + \sum_{i=1}^{N}(\sqrt{2c}x_i)(\sqrt{2c}y_i) + c^2\end{aligned}$$

上式共有 $N + \frac{N(N-1)}{2} + N + 1 = \frac{N^2+3N+2}{2} = \frac{(N+2)(N+1)}{2} = \binom{N+2}{2}$ 项. 其中，$\sum_{i=1}^{N}(x_i^2)(y_i^2)$ 中共有 N 项，$\sum_{i=2}^{N}\sum_{j=1}^{i-1}(\sqrt{2}x_ix_j)(\sqrt{2}y_iy_j)$ 共有 $\frac{N(N-1)}{2}$ 项，$\sum_{i=1}^{N}(\sqrt{2c}x_i)(\sqrt{2c}y_i)$ 共有 N 项，c^2 共有 1 项.
因此, 当 $d = 2$ 时,κ 把 N 维数据映射到了 $\binom{N+2}{2}$ 维空间中, 映射函数为

$$\Phi(\boldsymbol{x}) = (x_1^2, ..., x_N^2, \sqrt{2}x_2x_1, \sqrt{2}x_3x_1, \sqrt{2}x_3x_2, ..., \sqrt{2}x_Nx_{N-1}, \sqrt{2c}x_1, ..., \sqrt{2c}x_N, c)$$

$$\Phi(\boldsymbol{y}) = (y_1^2, ..., y_N^2, \sqrt{2}y_2y_1, \sqrt{2}y_3y_1, \sqrt{2}y_3y_2, ..., \sqrt{2}y_Ny_{N-1}, \sqrt{2c}y_1, ..., \sqrt{2c}y_N, c)$$

更一般地, 对 d 不加限制时,κ 把 N 维数据映射到 $\binom{N+d}{d}$ 维空间中.

[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

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) - 1\right). \tag{8}$$

然而 $\ell_{0/1}$ 数学性质不太好，它非凸、非连续，使得式 (8) 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 $\mathcal{L}_\Phi(h)=\underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\Phi(-yh(x))]$。定义 $L_\Phi(x,u)=\eta(x)\Phi(-u)+(1-\eta(x))\Phi(u)$，设 $h^*_\Phi(x)=\underset{u\in[-\infty,+\infty]}{\operatorname{argmin}}L_\Phi(x,u)$，$\mathcal{L}^*_\Phi=\mathcal{L}_\Phi(h^*_\Phi(x))$。

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

$$|h^*(x)|^s=\left|\eta(x)-\frac{1}{2}\right|^s\leq c^s\left[L_\Phi(x,0)-L_\Phi\left(x,h^*_\Phi(x)\right)\right] \tag{9}$$

则对于任何假设 h，有如下不等式成立

$$R(h)-R^*\leq 2c\left[\mathcal{L}_\Phi(h)-\mathcal{L}^*_\Phi\right]^{\frac{1}{s}} \tag{10}$$

(1) **[5 pts]** 请证明

$$\Phi\left(-2h^*(x)h(x)\right)\leq L_\Phi(x,h(x)) \tag{11}$$

(2) **[10 pts]** 请证明

$$R(h)-R^*\leq 2\underset{x\sim\mathcal{D}_x}{\mathbb{E}}\left[|h^*(x)|\,1_{h(x)h^*(x)\leq 0}\right] \tag{12}$$

提示：先证明

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

(3) **[10 pts]** 利用式 (11) 和式 (12) 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. 此处用于写解答 (中英文均可)

(1) *证明.*

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

□

(2) **证明.** 先证

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

因此

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

□

(3) **证明.** 由式 (12), *Jensen* 不等式, 式 (9), Φ 的非减性, 式 (11) 可知

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

$\square$

(4)

$$h_\Phi^*(x) = \underset{u\in[-\infty,+\infty]}{\operatorname{argmin}} L_\Phi(x,u) = 1_{\eta(x)>\frac{1}{2}} = 1_{h^*(x)>0}$$

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

$$L_\Phi(x, h_\Phi^*(x)) = {}_\Phi(x, 1_{\eta(x)>\frac{1}{2}}) = 1_{\eta(x)\le\frac{1}{2}} + 2(1-\eta(x))1_{\eta(x)>\frac{1}{2}} = 1_{h^*(x)\le 0} + 2(1-\eta(x))1_{h^*(x)>0}$$

将上述等式与 $s=1$, $c=\frac{1}{2}$ 代入 (9) 右边, 得 $\frac{1}{2}[1-1_{h^*(x)\le 0} - 2(1-\eta(x))1_{h^*(x)>0}]$.
我们考虑两种情况, 当 $h^*(x)\le 0$ 时, 右边 $=0$, 此时左边 $= h^*(x) \le 0 =$ 右边, 式 (9) 成立.
同样地, 当 $h^*(x)>0$ 时, 右边 $=\eta(x)-\frac{1}{2} = h^*(x)$, 此时左边 $= h^*(x) =$ 右边, 式 (9) 成立.
综上,

$$\forall h^*(x), h^*(x) \le \frac{1}{2}[1-1_{h^*(x)\le 0} - 2(1-\eta(x))1_{h^*(x)>0}]$$

成立.

[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

$$\hat{R}_{\text{LOO}}(\mathcal{A}) = \frac{1}{m}\sum_{i=1}^m 1_{h_{S-\{x_i\}}(x_i)\neq y_i}. \tag{13}$$

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

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_{\text{LOO}}(\mathcal{A})] = \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}[R(h_{S'})]. \tag{14}$$

(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

$$\mathbb{E}_{S\sim\mathcal{D}^m}[R(h_S)] \le \mathbb{E}_{S\sim\mathcal{D}^{m+1}}\left[\frac{N_{SV}(S)}{m+1}\right], \tag{15}$$

其中 $N_{SV}(S)$ 为 h_S 支持向量的个数。

Solution. 此处用于写解答 (中英文均可)

(1) **证明.** 由于 S 独立同分布,$\mathbb{E}_{S\sim\mathcal{D}^m}[1_{h_{S-x_i}(x_i)\neq y_i}]$ 不依赖于 i 的取值, 故

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}[1_{h_{S-x_i}(x_i)\neq y_i}]=\mathbb{E}_{S\sim\mathcal{D}^m}[1_{h_{S-x_1}(x_i)\neq y_i}]$$

因此证明如下:

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

□

(2) 如果 x 不是 h_S 的支持向量, 那么从 S 中去掉 x 并不影响 SVM 的解, 因此 $h_{S-x}=h_S$, 既然 h_S 可以将 x 正确分类, 那么 h_{S-x} 也可以将 x 正确分类.
该问题的逆否命题为: 若 h_{s-x} 将 x 错误分类, 则 x 必为 h_S 的支持向量.

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。

(3) **证明.** 因为若 h_{s-x} 将 x 错误分类, 则 x 必为 h_S 的支持向量, 即

$$\hat{R}_{\text{LOO}}(\mathcal{SVM})\leq\frac{N_{SV}(S)}{m+1}$$

两边同时取期望得

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

将式 (14) 带入左边, 得

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

□

[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)

(1) 代码见 **main.py**.

(2) 代码见 **main.py**.

(3) 代码见 **main.py**.

(4) 为与原代码区分, 我们将新设计的版本整合为 **NeuralNetwork_advanced** 类, 主要优化为可以自行调整模型隐藏层数和节点数, 最终我们将训练集随机打乱后, 按照 3:1 分为训练集和评估集, 不断调整模型隐藏层数, 节点数, 迭代次数以及学习率, 利用在评估集上得到较好结果的模型所选用的参数对整个训练集进行训练, 所得模型所对应于测试集的输出作为我们的输出结果.
输出文件见 **171860607_ypred.csv**.