题目

考虑具有连续权重的布尔型感知机的泛化误差在大 α 极限下的渐进形式 [1]

设定

考虑如下感知机模型

$$y = \mathrm{sign}\left(\frac{1}{\sqrt{N}}\mathbf{w}\cdot\mathbf{x}\right) \tag{1}$$

其中权重 $\mathbf{w}\in\mathbb{R}^N$ 满足球面约束 $\|\mathbf{w}\|^2 = N$，输入数据 $\mathbf{x}\in\mathbb{R}^N$。训练集 $\mathcal{D} = \left\{\left(\mathbf{x}^\mu, y_\star^\mu\right)\right\}_{\mu=1}^{P}$，标签

$$y_\star^\mu = \mathrm{sign}\left(\frac{1}{\sqrt{N}}\mathbf{w}_\star\cdot\mathbf{x}^\mu\right) \tag{2}$$

其中真实权重 $\mathbf{w}_\star\in\mathbb{R}^N$ 满足球面约束 $\|\mathbf{w}_\star\|^2 = N$，数据 $\mathbf{x}$ 满足高斯分布 $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_N)$。损失函数

$$\mathcal{L}(\mathbf{w}) = \sum_\mu \Theta\left(-y^\mu y_\star^\mu\right) \tag{3}$$

其中 Θ 为 Heaviside 阶跃函数。泛化误差

$$\varepsilon_{\mathrm{g}} = \mathbb{E}_{\mathbf{w}_\star,\mathbf{x}}\,\Theta\left(-y y_\star\right) \tag{4}$$

结论

定义数据量密度 $\alpha = P/N$，泛化误差可以表示为

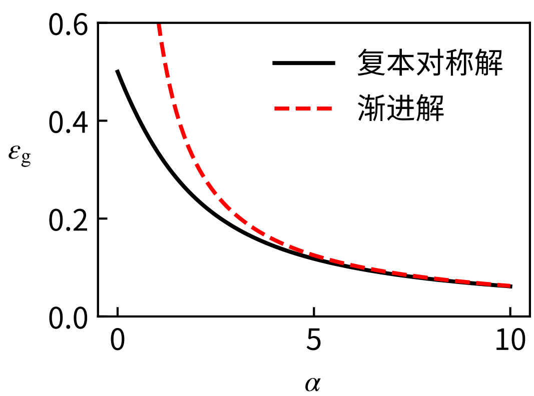
$$\varepsilon_{\mathrm{g}} = \frac{1}{\pi}\arccos r \tag{5}$$

其中 r 是如下方程的不动点

$$\frac{r}{\sqrt{1-r}} = \frac{\alpha}{\pi}\int \mathcal{D}x\,\frac{e^{-rx^2/2}}{\int_{\sqrt{r}x}^{\infty}\mathcal{D}z} \tag{6}$$

在 $\alpha\to\infty$ 时，泛化误差的渐进形式为

$$\varepsilon_{\mathrm{g}} \simeq \frac{0.625}{\alpha} \tag{7}$$



[1] 这个问题最早由 G. Györgyi 和 N. Tishby 在 1990 年得到复本对称解 [1]，H. S. Seung、H. Sompolinsky、N. Tishby 在 1992 年关于感知机的工作 [2] 和 H. Nishimori 在 2001 年的书 [3] 第八章对这个问题进行了介绍，有很高的参考价值

分析

1. 统计力学形式

配分函数

$$Z=\int \mathrm{d}\mathbf{w}\ \delta(N-\|\mathbf{w}\|^2)\exp\left[-\beta\sum_{\mu}\Theta\left(-y^{\mu}y_{\star}^{\mu}\right)\right] \tag{8}$$

注意到 $\lim_{\beta\to\infty}e^{-\beta\,\Theta(-x)}=\Theta(x)$，取零温极限，采用如下配分函数

$$Z=\int \mathrm{d}\mathbf{w}\ \delta(N-\|\mathbf{w}\|^2)\prod_{\mu}\Theta\left(y^{\mu}y_{\star}^{\mu}\right) \tag{9}$$

这对应于不考虑输出噪声的情况

自由能 $Nf=-\log Z$，复本技巧 $\mathbb{E}\log Z=\lim_{n\to0}\frac{1}{n}\log\mathbb{E}Z^n$，自由能密度写为

$$f=-\lim_{n\to0}\frac{1}{nN}\ \log\mathbb{E}_{\mathbf{w}_{\star},\{\mathbf{x}^{\mu}\}}Z^n \tag{10}$$

注意到 $\Theta\left[\mathrm{sign}(x_1)\,\mathrm{sign}(x_2)\right]=\Theta\left(x_1x_2\right)$，引入两个辅助场

$$u=\frac{1}{\sqrt{N}}\mathbf{w}_{\star}\cdot\mathbf{x}\qquad v^a=\frac{1}{\sqrt{N}}\mathbf{w}^a\cdot\mathbf{x} \tag{11}$$

自由能密度改写为

$$f=-\lim_{n\to0}\frac{1}{nN}\log\ \mathbb{E}_{u,\{v^a\}}\int\prod_a\frac{\mathrm{d}\mathbf{w}^a\mathrm{d}\hat{o}}{2\pi i}\exp\left(-\hat{o}N+\hat{o}\|\mathbf{w}^a\|^2\right)\left[\Theta\left(-uv^a\right)\right]^P \tag{12}$$

场变量满足联合分布 $(u,\{v^a\})\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})$，其中

$$\mathbf{\Sigma}_{uu}=\frac{\|\mathbf{w}_{\star}\|^2}{N}=1\qquad\mathbf{\Sigma}_{uv^a}=\frac{\mathbf{w}_{\star}\cdot\mathbf{w}^a}{N}\equiv r_a\qquad\mathbf{\Sigma}_{v^av^b}=\frac{\mathbf{w}^a\cdot\mathbf{w}^b}{N}\equiv q_{ab} \tag{13}$$

将序参量引入自由能中，得到

$$f=-\lim_{n\to0}\frac{1}{nN}\log\int\mathcal{D}O\mathcal{D}\hat{O}\exp\left[N\left(G_o+G_s+\alpha G_e\right)\right] \tag{14}$$

其中

$$\mathcal{D}O\mathcal{D}\hat{O}=\left(\prod_a\frac{\mathrm{d}r_a\mathrm{d}\hat{r}_a}{2\pi i/N}\right)\left(\prod_{a<b}\frac{\mathrm{d}q_{ab}\mathrm{d}\hat{q}_{ab}}{2\pi i/N}\right) \tag{15a}$$

$$G_o=-n\hat{o}-\sum_a\hat{r}_ar_a-\sum_{a<b}\hat{q}_{ab}q_{ab} \tag{15b}$$

$$G_s=\frac{1}{N}\log\int\prod_a\mathrm{d}\mathbf{w}^a\exp\left(n\hat{o}\|\mathbf{w}^a\|^2+\sum_a\hat{r}_a\mathbf{w}^a\cdot\mathbf{w}_{\star}+\sum_{a<b}\hat{q}_{ab}\mathbf{w}^a\cdot\mathbf{w}^b\right) \tag{15c}$$

$$G_e=\log\mathbb{E}_{u,\{v^a\}}\prod_a\Theta\left(-uv^a\right) \tag{15d}$$

在大 N 极限下，通过鞍点近似评估自由能密度

$$f=-\operatorname*{extr}_{O,\hat{O}}\left\{\mathcal{G}_o+\mathcal{G}_s+\alpha\mathcal{G}_e\right\} \tag{16}$$

其中 $\mathcal{G}=\lim_{n\to0}G/n$。

2. 复本对称假设

引入复本对称假设

$$r_a = r \qquad \hat{r}_a = \hat{r} \qquad q_{ab} = q \qquad \hat{q}_{ab} = \hat{q} \quad \text{for} \quad \forall a \neq b \tag{17}$$

式 (15b) 计算为

$$G_o = -n\hat{o} - nr\hat{r} - \frac{1}{2}n(n-1)q\hat{q} \tag{18}$$

式 (15c) 利用重参数化计算为 [2]

$$\begin{aligned} G_s &= \frac{1}{N}\log\int \mathcal{D}\mathbf{z}\left[\int \mathrm{d}\mathbf{w}\,\exp\left(\hat{r}\,\mathbf{w}_\star\cdot\mathbf{w} + \sqrt{\hat{q}}\,\mathbf{z}\cdot\mathbf{w} - \frac{1}{2}\left(\hat{q}-2\hat{o}\right)\|\mathbf{w}\|^2\right)\right]^n && \text{(19a)}\\ &= \frac{n}{N}\int \mathcal{D}\mathbf{z}\,\log\left(\sqrt{\frac{2\pi}{\hat{q}-2\hat{o}}}\right)^N \exp\left[\frac{\left(\hat{r}\,\mathbf{w}_\star + \sqrt{\hat{q}}\,\mathbf{z}\right)^2}{2\left(\hat{q}-2\hat{o}\right)}\right] && \text{(19b)}\\ &= \frac{n}{2}\log\frac{2\pi}{\hat{q}-2\hat{o}} + \frac{n}{2}\frac{\hat{q}+\hat{r}^2}{\hat{q}-2\hat{o}} && \text{(19c)} \end{aligned}$$

式 (15d) 利用重参数化计算为 [3]

$$\begin{aligned} G_e &= \log\mathbb{E}_{u,\{v^a\}}\,2\,\Theta\left(u\right)\prod_a \Theta\left(v^a\right) && \text{(21a)}\\ &= \log 2\int \mathcal{D}z_1\int \mathcal{D}z_2\,\Theta\left(\frac{r}{\sqrt{q}}z_1 + \sqrt{1-\frac{r^2}{q}}z_2\right)\left\{\int \mathcal{D}z_3\,\Theta\left(\sqrt{q}z_1 + \sqrt{1-q}z_3\right)\right\}^n && \text{(21b)}\\ &= 2n\int \mathcal{D}z_1\int_{-\frac{rz_1}{\sqrt{q-r^2}}}\mathcal{D}z_2\ \log\int_{-\sqrt{\frac{q}{1-q}}z_1}\mathcal{D}z_3 && \text{(21c)}\\ &= 2n\int_0^\infty \mathcal{D}\xi\int \mathcal{D}\zeta\ \log H\left(u\right) && \text{(21d)} \end{aligned}$$

[2]这个重参数化方案是利用 Hubbard-Stratonovich 变换得到的

$$\sum_{a<b}\hat{q}_{ab}\mathbf{w}^a\cdot\mathbf{w}^b = \frac{1}{2}\hat{q}\sum_{a\neq b}\mathbf{w}^a\cdot\mathbf{w}^b = \frac{1}{2}\hat{q}\left(\left(\sum_a \mathbf{w}^a\right)^2 - \sum_a\|\mathbf{w}^a\|^2\right) = -\frac{1}{2}n\hat{q}\|\mathbf{w}\|^2 + \log\int \mathcal{D}\mathbf{z}\,e^{n\sqrt{\hat{q}}\,\mathbf{w}^a\cdot\mathbf{z}}$$

[3]式 (21a) 利用了 $u>0,\ v^a>0$ 和 $u<0,\ v^a<0$ 的对称性
式 (21b) 中使用了如下重参数化方案

$$u = \frac{r}{\sqrt{q}}z_1 + \sqrt{1-\frac{r^2}{q}}\,z_2 \qquad v^a = \sqrt{q}\,z_1 + \sqrt{1-q}\,z_3^a$$

式 (21d) 中使用了如下重参数化方案

$$\zeta = \frac{r}{\sqrt{q}}z_2 - \sqrt{\frac{q-r^2}{q}}z_1 \qquad \xi = \sqrt{\frac{q-r^2}{q}}z_2 + \frac{r}{\sqrt{q}}z_1$$

定义了 $H(s) = \int_s^\infty \mathcal{D}x$ 以及

$$s = \frac{\sqrt{q-r^2}\,\zeta - r\xi}{\sqrt{1-q}} \tag{20}$$

因此

$$\mathcal{G}_o = -\hat{o} - r\hat{r} + \frac{1}{2}q\hat{q} \tag{22a}$$

$$\mathcal{G}_s = \frac{1}{2}\log\frac{2\pi}{\hat{q}-2\hat{o}} + \frac{\hat{q}+\hat{r}^2}{2\left(\hat{q}-2\hat{o}\right)} \tag{22b}$$

$$\mathcal{G}_e = -2\int_0^\infty \mathcal{D}\xi \int \mathcal{D}\zeta\ \log H\left(u\right) \tag{22c}$$

取无序平均后的自由能密度为

$$f = \operatorname{extr}\left\{\hat{o} + r\hat{r} - \frac{1}{2}q\hat{q} + \frac{1}{2}\log\left(\hat{q}-2\hat{o}\right) - \frac{\hat{q}+\hat{r}^2}{2\left(\hat{q}-2\hat{o}\right)} + 2\alpha\int_0^\infty \mathcal{D}\xi\int\mathcal{D}\zeta\ \log H\left(s\right)\right\} \tag{23}$$

3. 鞍点方程

令 $\partial_{\hat{o}}f = \partial_{\hat{r}}f = \partial_{\hat{q}}f = 0$，得到

$$1 = \frac{\hat{r}^2+2\hat{q}-2\hat{o}}{\left(q-2\hat{o}\right)^2} \qquad r = \frac{\hat{r}}{\hat{q}-2\hat{o}} \qquad q = \frac{\hat{q}+\hat{r}^2}{\left(\hat{q}-2\hat{o}\right)^2} \tag{24}$$

代入式 (23) 得到

$$f = \operatorname*{extr}_{r,q}\left\{-\frac{1}{2}\log\left(1-q\right) - \frac{1-r^2}{2\left(1-q\right)} + 2\alpha\int_0^\infty \mathcal{D}\xi\int\mathcal{D}\zeta\ \log H\left(s\right)\right\} \tag{25}$$

令 $\partial_r f = \partial_q f = 0$，得到

$$\frac{r}{1-q} = \frac{\alpha}{\sqrt{1-q}}\sqrt{\frac{2}{\pi}}\int_0^\infty \mathcal{D}\xi\int\mathcal{D}\zeta\ \frac{e^{-s^2/2}}{H(s)}\left(\frac{r\zeta}{\sqrt{q-r^2}}+\xi\right) \tag{26a}$$

$$-\frac{q+r^2}{2(1-q)^2} = \frac{\alpha}{\sqrt{1-q}}\sqrt{\frac{2}{\pi}}\int_0^\infty \mathcal{D}\xi\int\mathcal{D}\zeta\ \frac{e^{-s^2/2}}{H(s)}\left(\frac{\zeta}{2\sqrt{q-r^2}} + \frac{\sqrt{q-r^2}\zeta - r\xi}{2\left(1-q\right)}\right) \tag{26b}$$

式 (26a) 和 (26b) 的解满足 $q = r$ [4]，将其重写为 [5]

$$\frac{r}{\sqrt{1-r}} = \frac{\alpha}{\pi}\int\mathcal{D}x\ \frac{e^{-rx^2/2}}{\int_{\sqrt{r}x}^\infty \mathcal{D}z} \tag{27}$$

迭代式 (27) 至不动点，可以得到 r 的解

[4]这其实不需要计算也能想到，因为 $\mathbf{w}/\sqrt{N}$ 和 $\mathbf{w}_\star/\sqrt{N}$ 都是从球面 $\mathbb{S}^{N-1}$ 上随机均匀采样的，重叠序参量应该没有什么不同

[5]利用变量替换

$$\zeta \to x + \sqrt{\frac{r}{1-r}}y \qquad \xi \to y$$

4. 泛化误差

泛化误差 (4) 重写为 [6]

$$\varepsilon_g = \mathbb{E}_{u,v}\,\Theta(-uv) = 1 - 2\,\mathbb{E}_{u,v}\,\Theta(u)\,\Theta(v) = 1 - 2\int_0^\infty \mathcal{D}x \int_{-\frac{rx}{\sqrt{1-r^2}}} \mathcal{D}y = \frac{1}{\pi}\arccos r \tag{28}$$

考虑 $\alpha \to \infty$，$r \to 1$，令 $r = 1 - \delta$（$\delta \ll 1$），代入式 (27) 得

$$\frac{1}{\delta} = \frac{\alpha^2}{\pi^2}\left(\int \mathcal{D}x\,\frac{e^{-x^2/2}}{\int_x^\infty \mathcal{D}z}\right)^2 = \frac{\alpha^2}{1.926} \tag{29}$$

代入式 (28) 得到 [7]

$$\varepsilon_g = \frac{1}{\pi}\arccos(1-\delta) \simeq \frac{0.625}{\alpha} \tag{30}$$

References

[1] G. Györgyi and N. Tishby. Statistical theory of learning a rule. In W. K. Theumann and R. Köberle, editors, *Neural Networks and Spin Glasses*. World Scientific, Singapore, 1990.

[2] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

[3] H. Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. Clarendon Press, Oxford, 2001.

[6] 使用了如下重参数化方案

$$u = x \qquad v = rx + \sqrt{1-r^2}\,y$$

第二个等号利用了

$$\Theta(-uv) = \Theta(u)\,\Theta(v) + \Theta(-u)\,\Theta(-v) \qquad \text{和} \qquad \Theta(-u) = 1 - \Theta(u)$$

最后一步积分利用几何法得到

[7] 利用了

$$\arccos(x) \approx \sqrt{2(1-x)}$$