数学附录

1 一般机器学习为什么有偏？

一般机器学习算法的思路是：首先估计 g_0，再用 D 对 $Y-\hat{g}_0$ 做 OLS，得到 θ_0 的估计值：

$$\hat{\theta}_0=\left(\frac{1}{n}\sum_{i=1}^{n}D_i^2\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}D_i[Y_i-\hat{g}_0(X_i)]$$

代入 $Y_i=D_i\theta_0+g_0(X_i)+U_i$：

$$\hat{\theta}_0=\left(\frac{1}{n}\sum_{i=1}^{n}D_i^2\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}D_i[D_i\theta_0+g_0(X_i)+U_i-\hat{g}_0(X_i)]=\theta_0+\left(\frac{1}{n}\sum_{i=1}^{n}D_i^2\right)^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}D_iU_i+\frac{1}{n}\sum_{i=1}^{n}D_i[g_0(X_i)-\hat{g}_0(X_i)]\right\}$$

移项，代入 $D_i=m_0(X_i)+V_i$：

$$\sqrt{n}(\hat{\theta}_0-\theta_0)=\left(\frac{1}{n}\sum_{i=1}^{n}D_i^2\right)^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_iU_i+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m_0(X_i)[g_0(X_i)-\hat{g}_0(X_i)]+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i[g_0(X_i)-\hat{g}_0(X_i)]\right\}$$

由大数定律 (LLN)：

$$\frac{1}{n}\sum_{i=1}^{n}D_i^2\xrightarrow{p}E(D_i^2)$$

等价于：

$$\frac{1}{n}\sum_{i=1}^{n}D_i^2=E(D_i^2)+o_p(1)$$

所以：

$$\left(\frac{1}{n}\sum_{i=1}^{n}D_i^2\right)^{-1}=E(D_i^2)^{-1}+o_p(1)$$

带入 $\sqrt{n}(\hat{\theta}_0-\theta_0)$：

$$\sqrt{n}(\hat{\theta}_0-\theta_0)=E(D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_iU_i+E(D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}m_0(X_i)[g_0(X_i)-\hat{g}_0(X_i)]+E(D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i[g_0(X_i)-\hat{g}_0(X_i)]+o_p(1)$$

记：

$$\sqrt{n}(\hat{\theta}_0-\theta_0)=a+b+c+o_p(1)$$

对于 a，由于 U,V 独立且 $E[U|X]=0$：

$$E(D_iU_i)=0,\quad Var(D_iU_i)=E[(D_iU_i)^2]=E(D_i^2)E(U_i^2)$$

$$E(a)=0,\quad Var(a)=E(D_i^2)^{-2}\frac{1}{n}\sum_{i=1}^{n}E(D_i^2)E(U_i^2)$$

由大数定律 $\frac{1}{n}\sum_{i=1}^{n}E(D_i^2)E(U_i^2)$ 收敛，因此由中心极限定理 (CLT) 有：

$$a\xrightarrow{d}N(0,\Sigma)$$

对于 c，利用交叉拟合带来的独立性，且 $E[V|X]=0,\hat{g}_0\xrightarrow{p}g_0$：

$$E\{V_i[g_0(X_i)-\hat{g}_0(X_i)]\}=E(V_i)\cdot E[g_0(X_i)-\hat{g}_0(X_i)]=0\cdot 0=0$$

$$Var\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i[g_0(X_i)-\hat{g}_0(X_i)]\right\}=\frac{1}{n}\sum_{i=1}^{n}Var\{V_i[g_0(X_i)-\hat{g}_0(X_i)]\}$$

$$Var\{V_i[g_0(X_i)-\hat{g}_0(X_i)]\}=E(V_i^2)\cdot E\{[g_0(X_i)-\hat{g}_0(X_i)]^2\}$$

由大数定律，$\frac{1}{n}\sum_{i=1}^{n} E(V_i^2) \xrightarrow{p} E(V^2)$，且 $\|\hat{g}_0 - g_0\|_{L_2}^2 = o_p(1)$：

$$Var(\cdots) = E(V^2) \cdot o_p(1) = o_p(1) \to 0$$

所以：

$$c \xrightarrow{p} 0$$

对于 b：

$$b = E(D_i^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i)[g_0(X_i) - \hat{g}_0(X_i)]$$

具体地，$\hat{g}_0$ 到 g_0 的均方根误差收敛速度为 $n^{-\phi_g}$ 且 $\phi_g < 1/2$，$m_0(X_i) \neq 0$，因此 b 发散（$\sqrt{n}n^{-\phi_g} \to \infty$）。

所以，普通机器学习之所以有偏，因为 b 不收敛；$(g_0 - \hat{g}_0)$ 本身无法改成无偏，但可以改造 m_0 加快整体收敛速度，基于正交化方法就有了双重机器学习。

2 双重机器学习为什么无偏?

考虑部分线性模型：

$$E(Y_i|X_i) = E(D_i|X_i)\theta_0 + g_0(X_i) \implies l_0(X_i) = m_0(X_i)\theta_0 + g_0(X_i)$$

双重机器学习的思路是先估计 g_0, m_0（这个过程使用了两次机器学习，因此得名“双重”），然后使用 $\hat{V} = D - \hat{m}_0$ 对 $Y - \hat{l}_0$ 做 OLS，得到 θ_0 的估计值：

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i[Y_i - \hat{l}_0(X_i)]$$

代入 $Y_i = D_i\theta_0 + g_0(X_i) + U_i$：

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \left\{\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i[D_i\theta_0 + g_0(X_i) - \hat{l}_0(X_i) + U_i]\right\}$$

代入 $D_i = m_0(X_i) + V_i$，整理出 $\hat{\theta}_0 - \theta_0$：

$$\hat{\theta}_0 - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \left\{\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i[m_0(X_i)\theta_0 + g_0(X_i) - \hat{l}_0(X_i) + U_i]\right\}$$

即：

$$\hat{\theta}_0 - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \left\{\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i[l_0(X_i) - \hat{l}_0(X_i) + U_i]\right\}$$

因为 $\hat{V}_i = D_i - \hat{m}_0(X_i) = V_i + [m_0(X_i) - \hat{m}_0(X_i)]$，所以：

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} [V_i + m_0(X_i) - \hat{m}_0(X_i)][l_0(X_i) - \hat{l}_0(X_i) + U_i]\right\}$$

进一步展开上述乘积：

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \times$$

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_iU_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i[l_0(X_i) - \hat{l}_0(X_i)] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [m_0(X_i) - \hat{m}_0(X_i)]U_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [m_0(X_i) - \hat{m}_0(X_i)][l_0(X_i) - \hat{l}_0(X_i)]\right\}$$

由大数定律：

$$\left(\frac{1}{n}\sum_{i=1}^{n} \hat{V}_i^2\right)^{-1} \xrightarrow{p} E(V_i^2)^{-1}$$

记：

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^* + d^* + o_p(1)$$

同理：

$$a^* \xrightarrow{d} N(0, \Sigma^*)$$

$$b^* \xrightarrow{p} 0$$

$$c^* \xrightarrow{p} 0$$

对于 d^*：

$$d^* = E(V_i^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m_0(X_i) - \hat{m}_0(X_i)][l_0(X_i) - \hat{l}_0(X_i)]$$

这里由于前面的 m_0 变成了 $(m_0 - \hat{m}_0)$，l_0 是 m_0 和 g_0 的线性组合，$\widehat{m}_0$ 到 m_0 和 $\hat{g}_0$ 到 g_0 的收敛速度分别是 $n^{-\phi_m}$ 和 $n^{-\phi_g}$，乘积 $n^{-(\phi_m+\phi_m)}$，整体以更快的速度收敛，所以：

$$d^* \xrightarrow{p} 0$$

综上，由 Slutsky 定理：

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} N(0, \Sigma^*)$$

证明了 DDML 估计量的渐近无偏性与渐近正态性。