

b: as

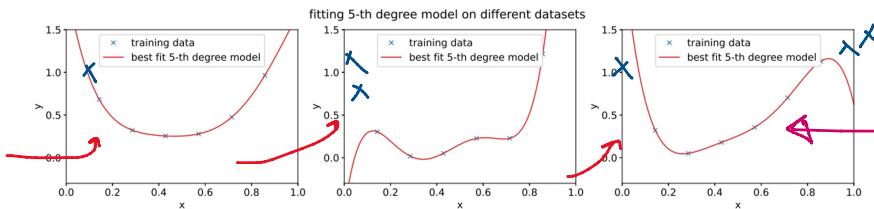
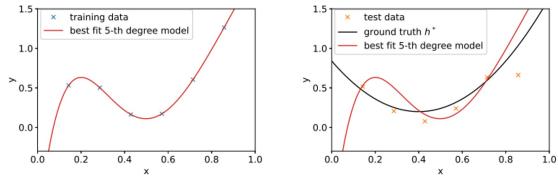


Figure 8.7: The best fit 5-th degree models on three different datasets generated from the same distribution behave quite differently, suggesting the existence of a large variance.

Yaser Abu Mostafa
Generalization
Bias / Voring

$$\text{I} \quad S = \{x^{(i)}, y^{(i)}\}_{i=1}^n$$

$$\text{II} \quad \hat{h}_S$$

$$\text{III} \quad h^*(x) + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

$$\boxed{\mathbb{E}_{S, \epsilon} [(y - h_S(x))^2]}$$

$$E_{A \sim A} = 0$$

$$\text{Id: } E((A+B)^2) = E(A^2) + E(B^2)$$

$$E_{s,\varepsilon}((y - h_s)^2) = E_{s,\varepsilon}((h^* + \varepsilon - h_s)^2)$$

$$\cancel{E[\varepsilon^2]} + E[(h^* - h_s)^2]$$

σ^2

$$h_{avg} = E_S(h_{S(\omega)})$$

average Model

$$\sigma^2 + E_S[(h^* - h_{avg})^2 + (h_{avg} - h_s)^2]$$

$$= \boxed{\sigma^2 + (h^* - h_{avg})^2 + E_S[(h_{avg} - h_s)^2]}$$

bias Variance s

۱. برای هر یک از موارد زیر، درستی یا نادرستی گزاره را با ذکر یک دلیل کوتاه مشخص کنید.

(آ) اگر مدل رگرسیون خطی با استفاده از ۸۰ درصد داده‌ها آموزش داده شود، نسبت به حالتی که از کل داده‌ها برای آموزش آن استفاده شود، بایاس کمتری خواهد داشت.

(ب) شرایط وجود دارد که تحت آن، هم بایاس و هم واریانس متناظر یک مدل، بالا باشد.

(ج) اگر دقت یک مدل روی داده‌های آموزشی خوب و روی داده‌های آزمون (تست) کم باشد، استفاده از مدلی پیچیده‌تر می‌تواند به رفع مشکل کمک کند.

۲. در مسئله رگرسیون، فرض کنید مقادیر y با استفاده از یک تابع چندجمله‌ای درجه ۵ از مقادیر x متغیر شان به دست آمداند. با در نظر گرفتن این امر، سه نوع مدل را در نظر گرفته‌ایم تا رابطه میان x و y را به صورت تخمینی مشخص کنیم. برای هر یک از مدل‌های زیر، با ذکر دلیل مختصر مشخص کنید پایاس و واریانس تخمین‌های متغیر شان روی داده‌ها (با توجه به مدل واقعی) را با عبارت‌های «کم» یا «زیاد» مشخص کنید.

- (آ) رگرسیون خطی
- (ب) چندجمله‌ای درجه ۵
- (ج) چندجمله‌ای درجه ۱۲

۳. در مسئله رگرسیون خطی با روش کمترین مربعات، می‌دانیم که می‌توان ضرایب مدل را از طریق تخمین‌گر $\hat{\theta}$ به دست آورد. تحت چه شرایطی، تخمین‌گر بیشینه درست‌نمایی نیز خواهد بود؟

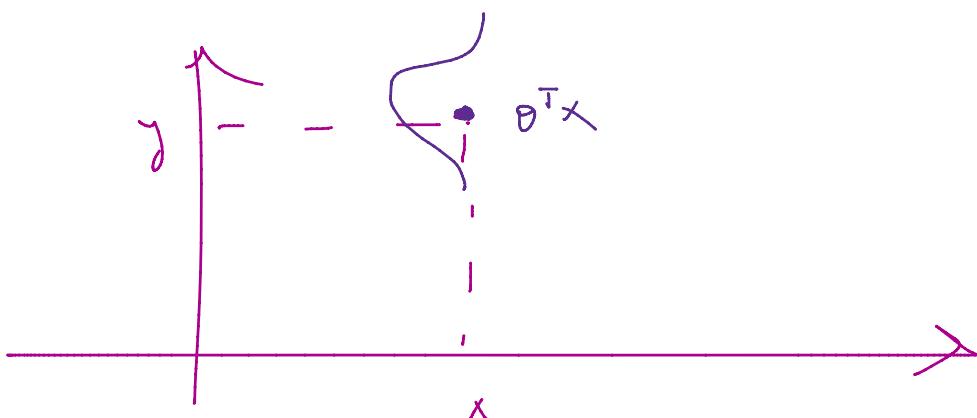
$$\text{MSE} \quad \hat{\theta} = \arg \max \prod P(y|x; \theta) \Rightarrow \hat{\theta} = \sum_i \log P(y|x_i; \theta)$$

$$\sum_i (\hat{y}_i - y_i)^2$$

$P(y|x) = \frac{ae^{b(y - \theta^T x)^2}}{1 + e^{b(y - \theta^T x)^2}}$

↗

$\partial \ell \quad \rightsquigarrow$



۴. رگرسیون خطی را در حالت ساده در نظر بگیرید که در آن قصد پیش‌بینی Y بر حسب X را داریم. ثابت کنید آماره‌ی R^2 برابر است با میاندور همبستگی X و Y . برای سادگی می‌توانید فرض کنید $\bar{x} = \bar{y} = 0$.

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{Var}(\beta_0 + \beta_1 x)}{\text{Var}(y)}$$

$$R^2 = \frac{\beta_1^2 \text{Var}(x)}{\text{Var}(y)}$$

$$\frac{\text{Cov}^2(x, y)}{\text{Var}(x) \text{Var}(y)}$$

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$y = \beta_0 + \beta_1 x$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 0$$

$$\sum_i (\beta_0 + \beta_1 x_i - y_i) = 0$$

$$n\beta_0 + \beta_1 \sum x_i - \sum y_i = 0$$

$$\beta_0 + \beta_1 x_i + \epsilon_i$$

۵. یک مسئله رگرسیون را در نظر بگیرید که در آن می خواهیم مقدار y را از روی یک مقدار حقیقی مانند x پیش‌بینی کنیم. تعدادی نمونه مانند $(x_i, y_i)_{i=1}^n$ داریم ($n \geq 3$). دو مدل زیر را برای مسئله رگرسیون خطی معمولی در نظر گرفته‌ایم.

$$y_i = w_0 + w_1 x_i + \epsilon_i \quad (1)$$

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i \quad (2)$$

توجه کنید که مقادیر ϵ_i مستقل و با توزیع یکسان (i.i.d.) هستند و از توزیع گوسی با میانگین صفر می‌آیند. به سوالات زیر پاسخ دهید.

(آ) در مدل (2)، رابطه‌ای برای تخمین w_2 به دست آورید. فرض کنید مقادیر w_0 و w_1 را می‌دانیم.

(ب) کدام یک از مدل‌ها دقت بیشتری را داده‌های آموزشی خواهد داشت؟ (با ذکر دلیل)

۱

$$J = \sum (y_i - (w_0 + w_1 x_i + w_2 x_i^2))^2$$

$$\frac{\partial J}{\partial w_2} = 0 \quad -2 \sum x_i^2 (y_i - (w_0 + w_1 x_i + w_2 x_i^2)) = 0$$

$$\sum x_i^2 y_i - \sum x_i^2 w_0 - \sum w_1 x_i$$

$$-w_2 \sum x_i^4 \Rightarrow w_2$$

۶. (رگرسیون خطی وزن‌دار): در مسئله رگرسیون خطی، قصد داریم به نمونه‌های آموزشی، وزن‌های متفاوتی نسبت دهیم. به بیان دقیق‌تر، می خواهیم مقدار $J(\theta)$ را کمینه کنیم که به صورت زیر تعریف می‌گردد:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m \underbrace{w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2}_{z_i}$$

(آ) نشان دهید ماتریس Z موجود است؛ به طوری که داریم:

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

(ب) با محاسبه $\nabla_{\theta} J(\theta)$ و برابر قرار دادن آن با صفر، مقدار θ -ای را که $J(\theta)$ را کمینه می‌کند، بیابید.

(توجه: در حالتی که همه وزن‌ها یکسان باشند، می‌دانیم $(X^T X)^{-1} X^T y = \theta^*$. جواب تابع برای این قسمت باید یک فرم بسته باشد که تابعی از X , W و y است).

(ج) فرض کنید مجموعه داده $\{(x^{(i)}, y^{(i)}) : i = 1, 2, \dots, m\}$ شامل m نمونه مستقل داده شده است. قصد داریم $y^{(i)}$ -ها را به گونه‌ای مدل کنیم که گویی از توزیع های شرطی با سطوح مختلفی از واریانس گرفته شده‌اند. به طور مشخص، فرض کنید داریم:

$$p(y^{(i)} | x^{(i)}, \theta) = \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

به بیان دیگر، $y^{(i)}$ از یک توزیع گوسی با میانگین $\theta^T x^{(i)}$ و واریانس $(\sigma^{(i)})^2$ می‌آید؛ $\sigma^{(i)}$ -ها ثابت هستند و مقدارشان مشخص است. نشان دهید که یافتن تخمین بیشینه درست‌نمایی برای θ ، معادل است با حل یک مسئله رگرسیون خطی وزن‌دار. به طور مشخص مقادیر $w^{(i)}$ -ها را بر حسب $\sigma^{(i)}$ -ها به دست آورید.

$$\boxed{\sqrt{T} \Delta .. \Delta \Delta \Delta \times \wedge \vee}$$

$$\boxed{\dots}$$

$$X^T A X = \sum_i \sum_j x_i A_{ij} x_j \rightarrow \sum_i x_i^2 A_i$$

$\underbrace{}_{\substack{i=j \\ i \neq j}}$ A_i
 $A_{ij} = 0$

$$W = \begin{bmatrix} \frac{1}{2} w_1 & & & \\ & \frac{1}{2} w_2 & \ddots & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} = \cdots []$$

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

$$\frac{\partial}{\partial \theta} (X^T A X) = \frac{(A + A^T)}{2A} \times$$

$$(\theta^T X - y^T) \omega (X \theta - y) \quad a^T b \quad \Sigma a_i b_i$$

$$= \theta^T \boxed{X^T W X} \theta - \underline{\theta^T X^T W y} - \underline{y^T W X \theta} + \underline{y^T W y}$$

$$\frac{2 X^T W X \theta}{2 X^T W X \theta} - 2 X^T W y - X^T W y$$

$$X^T W X \theta = X^T W y$$

$$\theta^* = (X^T W X)^{-1} (X^T W y)$$

$$TP(y^{(i)} | x^{(i)}) = \prod \frac{1}{\sqrt{2\pi}\sigma^{(i)}} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}}$$

$$= \prod \left[\frac{1}{\sqrt{2\pi}\sigma^{(i)}} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}} \right]$$

$$= \frac{1}{(2\pi)^{m/2}} \times \frac{1}{\prod \sigma_i} \times \left| e^{-\sum \frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}} \right|$$

$$\sum \frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}$$



$$\sigma_i = \sqrt{\omega_{ii}}$$