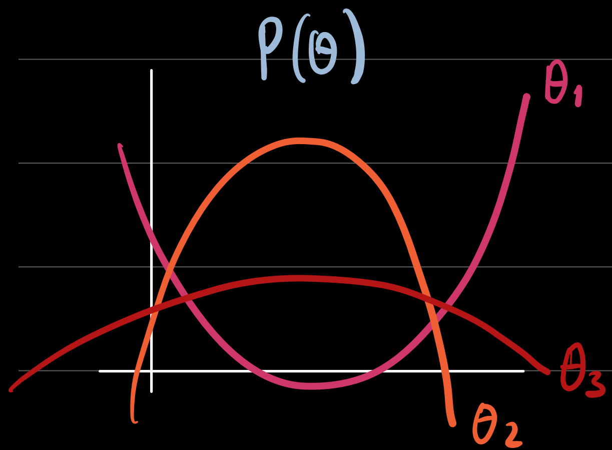
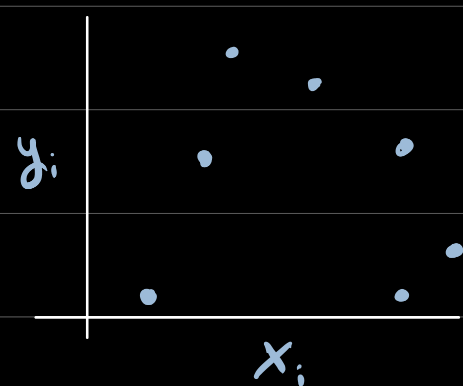
רגרסיה לינארית בייזיאנית - המשך

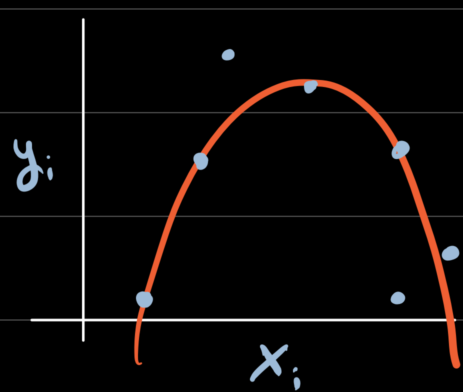
תזכורת: יש לנו פרדיקטור $y_\theta(x)=\sum_n \theta_n \cdot h_n(x)$

עם h_n פונקציות בסיס. ויש לנו פריור $P(\theta)$

לדוגמא: $h_1(x)=1\ ,\ h_2(x)=x\ ,\ h_3(x)=x^2$







משפט: אם $\theta\sim N(\mu_\theta,\Sigma_\theta)$ ו- $y|\theta\sim N(H\theta,\sigma^2 I)$

כאשר $H_{ij}=h_i(x_j)$

אז $\theta|D\sim N(\mu_{\theta|D},\Sigma_{\theta|D})$

כאשר

$$\mu_{\theta|D}=\left(\frac{1}{\sigma^2}H^TH+\Sigma_\theta^{-1}\right)^{-1}\left(\Sigma_\theta^{-1}\mu_\theta+\frac{1}{\sigma^2}H^Ty\right)$$

$$\Sigma_{\theta|D}=\left(\Sigma_\theta^{-1}+\frac{1}{\sigma^2}H^TH\right)^{-1}$$

מסקנה- $\theta^{MMSE}=\mu_{\theta|D}$ הוא הניבוי האופטימלי ל- θ ביחס ל- BMSE.

הערה - במקרה זה $\theta^{MAP}=\theta^{MMSE}=\mu_{\theta|D}$

הוכחה- מכיוון ש- $\theta|D\sim N(\mu_\theta,\Sigma_\theta)$, המקסימום מתקבל בתוחלת.

הערה- $\theta^{MLE}=\underset{\theta}{\arg\max}\,P(D|\theta)=(H^TH)^{-1}Hy$

משפט: אם $\theta \sim N(\mu_\theta, \Sigma_\theta)$ ו- $y|\theta \sim N(H\theta, \sigma^2 I)$

אזי

$$y(x_0)|D \sim N\left(h^T(x_0)\cdot \mu_{\theta|D},\ h^T(x_0)\Sigma_{\theta|D}\, h(x_0)\right)$$

כאשר $h^T(x_0) = h_1(x_0) + h_2(x_0) + \dots + h_n(x_0)$

הוכחה: מתכונות של גאוסיאנים והעתקות ליניאריות.

* ראו הוכחה בתרגול 1

מסקנה: הניבוי האופטימלי ל- $y(x_0)$ במובן של BMSE הוא $\mu_{y(x_0)|D} = h^T(x_0)\cdot \mu_{\theta|D}$

בחירת מודל בייזיאנית -
Bayesian Model Selection

כשאנחנו שואלים את עצמנו מהו המודל הנכון שמדל בעצם את הדאטא.

לדוגמא: $y^f(x)=\sum_{n=0}^{2}\theta_n X^n$, $y^g(x)=\sum_{n=0}^{10}\theta_n X_n$

כלומר יש לנו דאטא $D=\{X\}_{i=1}^{n}$ ואנחנו מחליטים אם הוא נוצר בעזרת פולינום מדרגה 2 או 10.

רעיון גרוע - נבחר מודל כך שה-MSE - $\sum_i (y_i - y_\theta(x_i))^2$ קטן ככל האפשר.

הסבר - אם משפחת הפונקציות $y^f(x)$ מוכלת במשפחת הפונקציות $y^g(x)$ אז בהכרח

$$MSE(y^{MLE,f}(x)) \geq MSE(y^{MLE,g}(x))$$

* y^{MLE} הוא ה-y שממזער את ה-MSE

Akaike Information Critirion (AIC)

עבור מודל רגרסיה עם n פונקציות בסיס $y|\theta \sim N(H\theta, \sigma^2 I)$

נבחר את המודל שממזער את

$$AIC = 2n + \frac{2}{\sigma^2}\sum_i (y_i - y_\theta(x_i))$$

ככל שמספר פונק' הבסיס יהיה גדול יותר, משלמים מחיר גדול יותר.

רעיון טוב- נגדיר מאורע H - משתנה בינארי

$H = \text{"}f\text{"}$ - הנתונים נדגמו מפונקציה $y^f(x)$

$H = \text{"}g\text{"}$ - הנתונים נדגמו מפונקציה $y^g(x)$

ונעניין בהתפלגות המותנית $P(H|D)$

$$P(H=\text{"}f\text{"}|D) = P(H=\text{"}f\text{"}) \cdot \underbrace{\boxed{P(D|H=\text{"}f\text{"})}}_{\text{evidence}} \cdot \frac{1}{P(D)}$$

כאשר נגדיר $P(D|H=\text{"}f\text{"}) := \int_\theta P(\theta|H=\text{"}f\text{"}) \cdot P(D|\theta, H=\text{"}f\text{"})$

משפט: נניח כי $\theta \sim N(\mu_\theta, \Sigma_\theta)$ ו- $y|\theta \sim N(H\theta, \sigma^2 I)$

ונניח שאנחנו יודעים את H. אזי מתקיים:

$$P(D) = \int_\theta P(\theta)\cdot P(D|\theta)\, d\theta = P(\hat{\theta}^{MAP}) P(D|\hat{\theta}^{MAP}) \cdot |\Sigma_{\theta|D}| \cdot (2\pi)^{\frac{n}{2}}$$

הערה 1) $P(D|\hat{\theta}^{MAP}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - y_{\hat{\theta}^{MAP}}(x_i)\right)^2\right)$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\cdot MSE\right)$$

הערה 2) הגורם $|\Sigma_{\theta|D}|$ מעדיף היפותזות שבהן יש הרבה θ-ות סבירות בהינתן הנתונים.
(זה גורם לכך שהוא מעדיף משפחות פרמטריות פשוטות. לאינטואיציה: אם $\Sigma_{\theta|D}$ הוא סקלר אז הביטוי יהיה קטן ככל שהשונות של $\theta|D$ קטנה יותר).

הוכחה: הראינו בשיעור שעבר

$$P(D)=\int_{\theta} exp\left(\left(\theta-\mu_{\theta|D}\right)^{T}\cdot\Sigma_{\theta|D}^{-1}\left(\theta-\mu_{\theta|D}\right)+C\right)d\theta$$

$$=exp(c)\cdot\int_{\theta} exp\left(\left(\theta-\mu_{\theta|D}\right)^{T}\cdot\Sigma_{\theta|D}^{-1}\left(\theta-\mu_{\theta|D}\right)\right)$$

(נמשיך בשבוע הבא).