Aujourd'hui : Motivations statistiques derrière Ridge et LASSO

→ Equilibre biais variance

→ Classification

→ Regression linéaire

$y(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$ $\quad \vec{x} \in \mathbb{R}^D$, $\{(x_i, t^{(i)})\}_{i=1}^N$

Pour rappel, dans le cadre de la minimisation RSS

On avait fait les hypothèses suivantes

$$t^{(i)} \sim \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} + \varepsilon^{(i)} \qquad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

$\varepsilon^{(i)}$ indépendants et identiquement distribués (i.i.d)

$$\varepsilon^{(i)} \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \rightarrow \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right)$$

$$p(\{t^{(i)}\}_{i=1}^N \mid \beta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right)$$

→ Estimateur de Maximum de vraisemblance

$$\beta_{MLE} = \underset{\beta}{\operatorname{argmax}}\ p(\{t^{(i)}\}_{i=1}^N \mid \beta)$$

En particulier, on a montré $\beta_{MLE} = \beta_{RSS}$

Théorème de Bayes

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \mid B)\, P(B)}{P(A)}$$

Estimateur de Maximum a Posteriori β_{MAP}

au lieu d'estimer le modèle en maximisant $p(\{t^{(i)}\}_{i=1}^{N} \mid \beta)$

on peut chercher le modèle qui maximise $p(\beta \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{N})$

En utilisant Bayes, on trouve

$$p(\underline{\beta} \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{N}) = \frac{p(\{t^{(i)}\}_{i=1}^{N} \mid \beta, \{x^{(i)}\}_{i=1}^{N})\, p(\beta)}{p(\{t^{(i)}\}_{i=1}^{N})}$$

$p(\{t^{(i)}\}_{i=1}^N) = \sum_{\beta_k} p(\{t^{(i)}\}_{i=1}^N \mid \beta_k)\, p(\beta_k)$ → indépendant de β

$$\beta_{MAP} = \text{argmax} \; p(\{t^{(i)}\}_{i=1}^N \mid \beta, \{x^{(i)}\}_{i=1}^N)\, p(\beta)$$

Exemples de choix pour $p(\beta)$: → $p(\beta) \sim$ Uniforme

$\beta_{MAP} = \beta_{MLE}$

→ $p(\beta) \sim$ Gaussienne

$$p(\beta) \sim \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\xi} \exp\left(-\frac{\beta_j^2}{2\xi^2}\right)$$

β_j indépendant et indépend. distribué selon une loi Normale

→ Pour un a priori Gaussien sur les β_j l'estimateur β_{MAP} se récrit

$$\beta_{MAP} = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right) \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\xi} \exp\left(-\frac{\beta_j^2}{2\xi^2}\right)$$

On commence par appliquer le logarithme ce qui donne

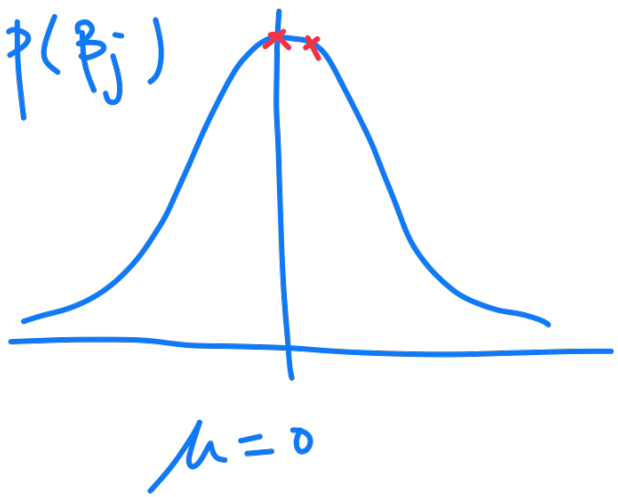
$$\beta_{MAP} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2} + \sum_{j=1}^{D} \log\left(\frac{1}{\sqrt{2\pi}\xi}\right) - \frac{\beta_j^2}{2\xi^2}$$

indépendants de β

$$\beta_{MAP} = \arg\max_{\beta} \; -\sum_{i=1}^{N} \frac{\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2}{2\sigma^2} - \sum_{j=1}^{D} \frac{\beta_j^2}{2\xi^2}$$

$$= \arg\min_{\beta} \; \sum_{i=1}^{N} \frac{\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{\beta_j^2}{2\xi^2}$$

$\beta_{RIDGE} = \beta_{MAP}$ avec un a priori Gaussien sur les β_j



→ Pour favoriser davantage l'annulation des coefficients on peut choisir un a priori de type Laplace

$$p(\beta_j) = \frac{1}{2\lambda} \exp\left(-\frac{|\beta_j|}{\lambda}\right)$$

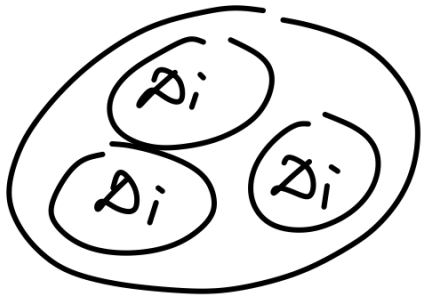
$$\hat{\beta}_{MAP} = \arg\max_{\beta} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - h_\beta(x^{(i)}))^2}{2\sigma^2}\right) \prod_{j=1}^{D} \frac{1}{2\xi} \exp\left(-\frac{|\beta_j|}{\xi}\right)$$

en appliquant le log et en transformant la maximisation en minimisation, on retrouve

$$\beta_{MAP} = \operatorname{argmin} \sum_{i=1}^{N} \frac{\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{|\beta_j|}{\xi}$$

Equilibre biais - variance

$(x, t(x))$ donné test



$\mathcal{D} = \{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$

Erreur quadratique moyenne

$$MSE = \mathop{\mathbb{E}}_{\mathcal{D}_i} \{ (h_{\mathcal{D}_i}(x) - t(x))^2 \}$$

$$= \mathop{\mathbb{E}}_{\mathcal{D}_i} \{ (h_{\mathcal{D}_i}(x) - \mathop{\mathbb{E}}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) + \mathop{\mathbb{E}}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) - t(x))^2 \}$$

En appliquant le carré, on obtient

déterministe

$$MSE(x) = \mathbb{E}_{\mathcal{D}_i}\left\{ \left(h_{\mathcal{D}_i}(x) - \mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) \right)^2 \right\} + \mathbb{E}_{\mathcal{D}_i}\left\{ \left(\mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) - t(x) \right)^2 \right\}$$

$$+ 2\,\mathbb{E}_{\mathcal{D}_i}\left\{ \left(h_{\mathcal{D}_i}(x) - \mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) \right) \left(\mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) - t(x) \right) \right\}$$

$= 0$

$$MSE(x) = \underbrace{\mathbb{E}_{\mathcal{D}}\left\{ \left(h_{\mathcal{D}_i}(x) - \mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) \right)^2 \right\}}_{\text{Variance}} + \underbrace{\left(\mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) - t(x) \right)^2}_{\text{bias}^2}$$