Maximum A Posteriori Estimation (MAP)

$$f_{Bayes}(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}]$$

$$= \int_y y \, p_{Y|\vec{X}}(y|\vec{x}) \, dy$$

If $p_{Y|\vec{X}}$ is a function of some parameter $w \in W$ then we just need to estimate that parameter

MLE: $\arg\max_{w \in W} \underbrace{p(D|w)}_{= \prod_{i=1}^{n} p(z_i|w)}$

"find w that maximizes the likelihood of the data"

MAP: $\arg\max_{w \in W} \underbrace{p(w|D)}_{= \text{"posterior"}}$

"find w that is the most likely given the data"

$$w_{MAP} = \arg\max_{w \in W} p(w|D)$$

$$= \arg\max_{w \in W} \frac{p(w, D)}{p(D)}$$

prod rule ↘

$$= \arg\max_{w \in W} \frac{p(D|w)\,p(w)}{p(D)}$$

$$= \arg\max_{w \in W} \underbrace{p(D|w)}_{\text{likelihood}}\,\underbrace{p(w)}_{\text{prior}}$$

MAP Basics

Ex: $Z_i \sim \mathcal{N}(\mu^*, 1)$ is the i-th persons height

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(z-\mu^*)^2}{2}\right)$$

$$\mu \sim \mathcal{N}(160, \sigma^2), \quad p(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\mu-160)^2}{2\sigma^2}\right)$$

$$\mu_{MAP} = \arg\max_{\mu \in \mathbb{R}} p(D|\mu)\,p(\mu)$$

$$= \arg\min_{\mu \in \mathbb{R}} -\log\left(p(D|\mu)\,p(\mu)\right) \qquad \log(ab) = \log(a) + \log(b)$$

$$= \arg\min_{\mu \in \mathbb{R}} \left[-\log\left(p(D|\mu)\right) - \log\left(p(\mu)\right)\right]$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \frac{(z_i - \mu)^2}{2} - \log\left(p(\mu)\right) \right]$$

$$\log(p(\mu)) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\mu - 160)^2}{2\sigma^2} \right) \right)$$

$$= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log\left(\exp\left(\frac{-(\mu - 160)^2}{2\sigma^2} \right) \right)$$

$$= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(\mu - 160)^2}{2\sigma^2}$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \frac{(z_i - \mu)^2}{2} - \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \frac{(\mu - 160)^2}{2\sigma^2} \right]$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \underbrace{\left[\sum_{i=1}^{n} \frac{(z_i - \mu)^2}{2} + \frac{(\mu - 160)^2}{2\sigma^2} \right]}_{g(\mu)}$$

$$\frac{dg}{d\mu}(\mu) = -\sum_{i=1}^{n} (z_i - \mu) + \frac{(\mu - 160)}{\sigma^2} = 0$$

$$\Rightarrow -\sum_{i=1}^{n} z_i + n\mu + \frac{\mu}{\sigma^2} - \frac{160}{\sigma^2} = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_i + \frac{160}{\sigma^2} = \left(n + \frac{1}{\sigma^2} \right) \mu$$

$$\Rightarrow \mu_{MAP} = \mu = \frac{\sum_{i=1}^{n} z_i + \frac{160}{\sigma^2}}{n + \frac{1}{\sigma^2}} \quad , \mu_{MLE} = \frac{\sum_{i=1}^{n} z_i}{n}$$

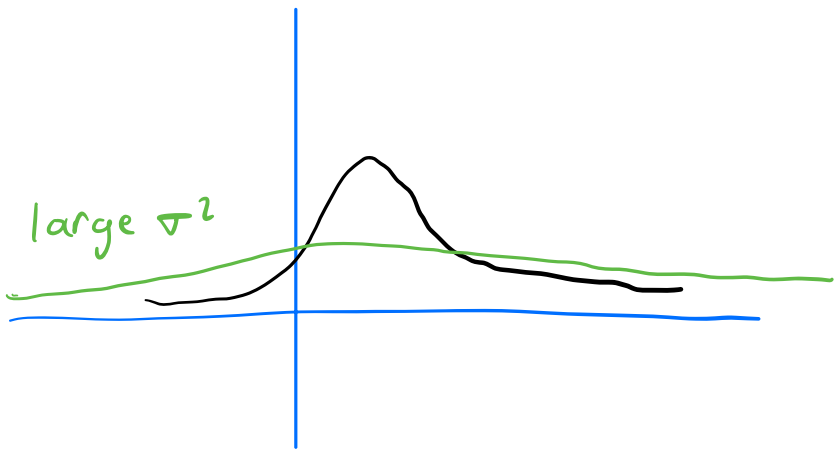
if n is large then $\mu_{MAP} \approx \frac{\sum_{i=1}^{n} z_i}{n} = \mu_{MLE}$

if σ^2 is small then $\mu_{MAP} \simeq \frac{\frac{160}{\sigma^2}}{\frac{1}{\sigma^2}} = 160$

if σ^2 is large then $\mu_{MAP} \approx \mu_{MLE}$

if $p(\mu) \simeq C \in \mathbb{R}$

then $\mu_{MAP} = \mu_{MLE}$



Estimating $\vec{w}$ for $P_{Y|\vec{X}}$

$D = ((\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, $\mathbb{P}_D$, p_D

$(\vec{X}_i, Y_i)$ are i.i.d with $\mathbb{P}_{\vec{X},Y}$ and $p_{\vec{X},Y}$

Assume $Y_i | \vec{X}_i = \vec{x}_i \sim \mathcal{N}(\vec{x}_i^T \vec{w}^*, 1)$

$$p_{Y|\vec{X}=\vec{x}}(y|\vec{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y - \vec{x}^T \vec{w}^*)^2}{2}\right)$$

Assume $w_j \sim \mathcal{N}(0, 1/\lambda)$ are i.i.d. for $j \in \{1, \ldots, d\}$

and $w_0 \sim \mathcal{N}(0, \sigma^2)$ for very large σ

$\approx \text{Uniform}(-a, a)$ for large a

w_0 is independent of w_j for all $j \in \{1, \ldots, d\}$

Calculating $\vec{w}_{MAP}$:

$$\vec{w}_{MAP} = \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \; p(\vec{w} \mid D)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \; p(D \mid \vec{w}) \, p(\vec{w})$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \; -\log\left(p(D \mid \vec{w}) \, p(\vec{w})\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left[-\log\left(p(D \mid \vec{w})\right) - \log\left(p(\vec{w})\right)\right]$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{w})^2}{2} - \log\left(p(\vec{w})\right)\right]$$

$$\log\left(p(\vec{w})\right) = \log\left(p(w_0, w_1, \ldots, w_d)\right)$$

$$\overset{\text{Independence}}{=} \log\left(p(w_0) \, p(w_1) \cdots p(w_d)\right)$$

$$= \log\left(p(w_0) \prod_{j=1}^{d} p(w_j)\right)$$

$$= \log\left(p(w_0)\right) + \sum_{j=1}^{d} \log\left(p(w_j)\right)$$

$$= \log\left(p(w_0)\right) + \sum_{j=1}^{d} \log\left(\frac{1}{\sqrt{2\pi/\lambda}} \exp\left(-\frac{w_j^2}{2/\lambda}\right)\right)$$

$$= \log\left(p(w_0)\right) + \sum_{j=1}^{d} \left[\log\left(\frac{1}{\sqrt{2\pi/\lambda}}\right) - \frac{w_j^2}{2/\lambda}\right]$$

$$= \log(p(w_0)) + d\log\left(\frac{1}{\sqrt{2\pi/\lambda}}\right) - \sum_{j=1}^{d} \frac{\lambda w_j^2}{2}$$

$$= \operatorname{argmin}_{\vec{w}\in\mathbb{R}^{d+1}} \left[\sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{w})^2}{2} - \log(p(w_0)) - d\log\left(\frac{1}{\sqrt{2\pi/\lambda}}\right) + \sum_{j=1}^{d} \frac{\lambda w_j^2}{2} \right]$$

$$= \operatorname{argmin}_{\vec{w}\in\mathbb{R}^{d+1}} \left[\underbrace{\sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{w})^2}{2}}_{= \frac{n}{2}\hat{L}} + \underbrace{\frac{\lambda}{2} \sum_{j=1}^{d} w_j^2}_{\text{almost regularizer}} \right]$$

$$= \operatorname{argmin}_{\vec{w}\in\mathbb{R}^{d+1}} \left[\frac{n}{2} \hat{L}_\lambda(\vec{w}) \right] \qquad \frac{1}{n}\sum_{i=1}^{n} (y_i - \vec{x}_i' \vec{w})^2 + \frac{\lambda}{n}\sum_{j=1}^{d} w_j^2$$

$$= \operatorname{argmin}_{\vec{w}\in\mathbb{R}^{d+1}} \left[\hat{L}_\lambda(\vec{w}) \right] = \vec{w}_{MAP} = \hat{\vec{w}}_\lambda$$

$$f_{Bayes}(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}]$$

$$= \int_y y \, p_{Y|\vec{x}}(y|\vec{x}) \, dy$$

$$\approx \int_y y \, p(y|\vec{x}, \vec{w}_{MAP}) \, dy$$

$$= \mathbb{E}[Y' \mid \vec{X} = \vec{x}] \quad \text{where } Y' \mid \vec{X} = \vec{x} \sim \mathcal{N}(\vec{x}^T \vec{w}_{MAP}, 1)$$

$$= \vec{x}^T \vec{w}_{MAP}$$

$$= \vec{x}^T \hat{\vec{w}}_\lambda = \hat{f}$$

Assume $W_j \sim \text{Laplace}(0, 1/\lambda)$ are i.i.d for $j \in \{1, \ldots, d\}$

and $W_0 \sim \text{Laplace}(0, b)$ for very large b

$\simeq \text{Uniform}(-a, a)$ for large a

W_0 is independent of W_j for all $j \in \{1, \ldots, d\}$

$$\vec{w}_{MAP} = \arg\max_{\vec{w} \in \mathbb{R}^{d+1}} p(\vec{w} \mid D)$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} \left[\sum_{i=1}^{n} \frac{(y_i - \vec{x}_i^T \vec{w})^2}{2} + \lambda \sum_{j=1}^{d} |w_j| \right]$$

Lasso regression