11/08/25

MAP (Maximum a posteriori) Estimate

- linear regression → MLE
- MAP → Adding regularizer to the loss function

$\ell(y, \hat{y})$ y = ground truth , $\hat{y}$ = prediction

x = feature vector

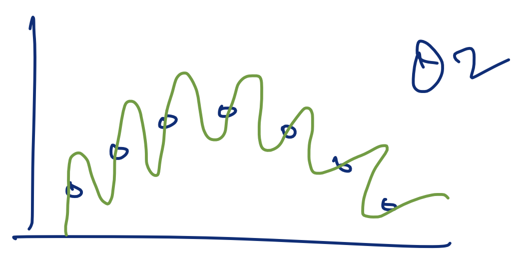
$\ell(\theta; x, y) \equiv \ell(y, f_\theta(x))$

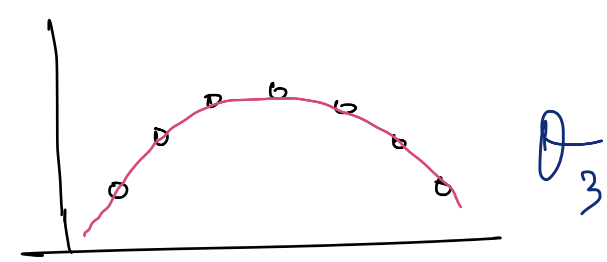
$\mathcal{D} = \{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$

$$\arg\max_\theta \sum_{i=1}^{N} \log p_\theta(y_i | x_i) = \arg\min_\theta - \sum \log p_\theta(y_i | x_i)$$

negative log likelihood

MLE





$p(\theta_3) > p(\theta_2) > p(\theta_1)$

There are infinitely many θ that are the MLE

- MLE has no notion of which of these parameters might be better than other.
- Prior belief about the data points.

$p(y|x) \sim$

$y|x \sim RV$

$x \sim RV$

- Treat the parameters θ as a RV

$p(\theta)$ – prior distribution that encodes which functions are more likely than others.

What $P(\theta)$ makes smooth f_θ more likely and non-smooth f_θ less likely?

$\rightarrow \quad \theta, \quad P(\theta)$

Roughly speaking, as the magnitue of θ increase, $P(\theta)$ shoud decrease

$$P(\theta) = \mathcal{N}(\theta; 0, \sigma^2 I) \quad \text{— MVG}$$

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$$

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)\, P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D} \mid \theta)$ → likelihood

$P(\theta)$ → prior

$P(\mathcal{D})$ →

$$P(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} P(x_i)\, P(y_i \mid x_i, \theta) \qquad P_\theta(y_i \mid x_i)$$

$$\log P(\theta \mid \mathcal{D}) = \sum_{i=1}^{N} \{\cancel{\log p(x_i)} + \log p(y_i \mid x_i, \theta)\} + \log P(\theta)$$

arg max

$$= \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta) + \log P(\theta)$$

$$\arg\max_\theta = \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta) + \log P(\theta)$$

$$\arg\min_\theta \sum -\log P(y_i \mid x_i, \theta) - \log P(\theta)$$

MVG
$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$\mu = 0$

$$= \frac{1}{(2\pi)^{d/2} (\sigma^2)^{d/2}} \exp\left(-\frac{1}{2} x^T (\sigma^2 I)^{-1} x\right)$$

isotropic Gaussian $\Sigma = \sigma^2 I_{d \times d}$

$$= \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2} x^T \frac{1}{\sigma^2} I x\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} x^T x\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right)$$

$$P(\theta) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\theta\|_2^2}{2\sigma^2}\right) =$$

$\hookrightarrow \arg\min_{\theta} \sum_{i=1}^{N} -\log p(y_i | x_i; \theta) + \dfrac{\|\theta\|_2^2}{2\sigma^2}$ — Regularization

$$\underbrace{(x\theta - y)^2 + \lambda \|\theta\|_2^2}_{} \quad \lambda = \text{param.}$$

$$\nabla_\theta = \quad (\theta^T x^T x\theta - 2y^T x\theta + y^T y) + \lambda \|\theta\|_2^2$$

$$= (\theta^T x^T x\theta - 2y^T x\theta + \lambda \theta^T \theta)$$

$$= \theta^T (x^T x + \lambda I)\theta - 2y^T x\theta$$

$$\nabla_\theta = 2\theta(x^T x + \lambda I) - 2X^T y = 0$$

$$\theta_{MAP} = (x^T x + \lambda I)^{-1} x^T y \quad \Big| \quad \theta_{MLE} = (x^T x)^{-1} x^T y$$

$$\arg\min \|x\theta - y\|_2^2 + \lambda \|\theta\|_2^2$$

λ = hyper parameter

L2 regularization

Ridge Reg.

LASSO Regression / L_1 Regularizer

$$P(\theta_i) = \frac{1}{2b} \exp\left(-\frac{|\theta_i|}{b}\right) \qquad P(\theta_i) \sim \text{laplace}(\theta_i, 0, b)$$

$d = M$

$$\log P(\theta_i) = -\frac{1}{b}|\theta_i| + C$$

$$\arg\min \|X\theta - Y\|_2^2 + \underbrace{\frac{1}{b}}_{\lambda} \underbrace{\sum_{i=1}^{d} \theta_i}_{\|\theta\|_1}$$

$$\|X\theta - Y\|_2^2 + \lambda \|\theta\|_1$$