3. Bayesian Interpretation of Regularization

⇒ In Bayesian statistics, almost every quantity is a random variable, which can either be observed or unobserved.

θ → unobserved random variable

x & y → observed random variable

⇒ The joint distribution of all the random variable is also called the model $(P(x, y, \theta))$.

⇒ Every unknown quantity can be estimated by conditioning the model on all the observed quantities.

↳ Such a conditional distribution is known as Posterior distribution.

⇒ A consequence of this approach is that, we are required to endow our model parameters, $P(\theta)$ with a prior distribution.

↓

{The prior probabilities are to be assigned before we see the data}

⇒ Estimating the mode of the posterior distribution is also called maximum a posteriori estimate (MAP).

$$\boxed{\theta_{MAP} = \arg\max_{\theta} P(\theta \mid x, y)}$$

⇒ On contrary maximum likelihood estimate (MLE)

$$\boxed{\theta_{MLE} = \arg\max_{\theta} P(y \mid x, \theta)}$$

(a) $\theta_{MAP} = \arg\max_{\theta} P(y|x;\theta)\, P(\theta)$ {To prove}

$\theta_{MAP} = \arg\max_{\theta} P(\theta|x,y)$

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

$= \arg\max_{\theta} \frac{P(y|x,\theta)\, P(\theta|x)}{P(y)}$

$= \arg\max_{\theta} P(y|x,\theta)\, P(\theta|x)$ {P(y) is not a function of θ}

$\boxed{\theta_{MAP} = \arg\max_{\theta} P(y|x,\theta)\, P(\theta)}$ {Assuming $P(\theta) = P(\theta|x)$}

(b) * L_2 regularization penalizes the L_2 norm of the parameters while minimizing the loss

⇒ Show that MAP estimation with a zero-mean Gaussian prior over θ (i.e. $\theta \sim N(0, \eta^2 I)$), is equivalent to applying L_2 regularization with MLE estimation.

$\theta_{MAP} = \arg\min_{\theta} -\log P(y|x,\theta) + \lambda \|\theta\|_2^2$

$P(\theta) = \frac{1}{\sqrt{(2\pi)^n \eta^{2n}}} \exp\left(-\frac{1}{2} \theta^T (\eta^2 I)^{-1} \theta\right)$

$= \frac{1}{\sqrt{(2\pi\eta^2)^n}} \exp\left(-\frac{1}{2\eta^2} \theta^T \theta\right)$

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

$$\theta_{MAP} = \underset{\theta}{\arg\max} \log\left(P(y|x,\theta)\, P(\theta)\right)$$

$\{$ as $\log(x)$ is monotonic increasing function of x $\}$

$$\theta_{MAP} = \underset{\theta}{\arg\max}\left[\log(P(y|x,\theta)) + \underbrace{\log(P(\theta))}_{\log\left(\frac{1}{(2\pi\eta^2)^{n/2}}\right) - \frac{1}{2\eta^2}\|\theta\|_2^2}\right]$$

$$\boxed{\theta_{MAP} = \underset{\theta}{\arg\min} -\log P(y|x,\theta) + \frac{1}{2\eta^2}\|\theta\|_2^2}$$

$$\boxed{\text{Where } \lambda = \frac{1}{2\eta^2}}$$

(c) $\Rightarrow$ Consider a linear regression model given by

$$y = \theta^T x + \epsilon \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

$$\theta \sim N(0, \eta^2 I) \quad \{\text{Gaussian prior}\}$$

$\Rightarrow$ Let X be the design matrix:

$$X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(m)T} \end{bmatrix}$$

$\Rightarrow$ & $$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

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

$$-\log P(Y|X,\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)},\theta) \quad \{\text{IID Assumption}\}$$

$$= \sum_{i=1}^{m} -\log P(y^{(i)}|x^{(i)},\theta)$$

$$\downarrow$$

$$-\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}$$

$$\theta_{MAP} = \arg\min_{\theta}\left(\sum_{i=1}^{m} \frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2} + \frac{1}{2\eta^2}\|\theta\|_2^2\right)$$

$$\| \\ J$$

$$\boxed{J = \frac{1}{2\eta^2}\theta^T\theta + \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^{(i)}-\theta^T x^{(i)})^2} \quad \{\text{Cost function}\}$$

$\Rightarrow$ For the value of θ for which $J(\theta)$ is minimum can be found by

$$\nabla_\theta J(\theta) = 0$$

$$\nabla_\theta J(\theta) = \frac{\theta}{\eta^2} + \frac{1}{2\sigma^2}\sum_{i=1}^{m} 2(y^{(i)}-\theta^T x^{(i)})(-x^{(i)})$$

$$\downarrow$$

$$-\frac{1}{\sigma^2}\sum_{i=1}^{m} y^{(i)}x^{(i)} - (\theta^T x^{(i)})x^{(i)}$$

$$-\frac{X^TY}{\sigma^2} \leftarrow -\frac{1}{\sigma^2}\sum_{i=1}^{m} y^{(i)}x^{(i)}$$

$$+ \frac{1}{\sigma^2}\sum_{i=1}^{m}(\theta^T x^{(i)})x^{(i)} \rightarrow \frac{X^TX\theta}{\sigma^2}$$

$$\nabla_\theta J(\theta) = -\frac{X^T Y}{\sigma^2} + \frac{X^T X \theta}{\sigma^2} + \frac{\theta}{\eta^2} = 0$$

$$\Rightarrow \frac{X^T X \theta}{\sigma^2} - \frac{X^T Y}{\sigma^2} + \frac{\theta}{\eta^2} = 0$$

$$\left(\frac{X^T X}{\sigma^2} + \frac{I}{\eta^2}\right)\theta = \frac{X^T Y}{\sigma^2}$$

$$\boxed{\theta_{MAP} = \left(\frac{X^T X}{\sigma^2} + \frac{I}{\eta^2}\right)^{-1} \left(\frac{X^T Y}{\sigma^2}\right)}$$

(d) ⇒ Consider Laplace distribution, whose density is given by:

$$\boxed{f_L(z|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|z-\mu|}{b}\right)}$$

⇒ Consider linear regression model given by

$$y = x^T\theta + \epsilon \quad \text{where } \epsilon \sim N(0,\sigma^2)$$

⇒ Assume a Laplace prior on this model where $\theta \sim \mathcal{L}(0, bI)$

⇒ Show that θ_{MAP} in this case is equivalent to the solution of linear regression with L_1 regularization whose loss function is specified as

$$\boxed{J(\theta) = \|X\theta - Y\|_2^2 + \gamma\|\theta\|_1}$$

$$\theta_{MAP} = \arg\max_\theta P(y|x,\theta)P(\theta)$$

$$= \arg\min_\theta -\log P(y|X,\theta) - \log(P(\theta))$$

$$-\log\left(\frac{1}{2b}\exp\left(-\frac{|\theta|}{b}\right)\right)$$

$$-\log\left(\frac{1}{2b}\right) + \frac{1}{b}\|\theta\|_1$$

$$J = -\log P(Y|X,\theta) + \frac{1}{b}\|\theta\|_1$$

$$J = \|X\theta - Y\|_2^2 + \gamma\|\theta\|_1 \quad \text{where } \boxed{\gamma = 1/b}$$

Rigid regression

→ Linear regression with L_2 regularization is also commonly called Rigid regression.

Lasso regression

→ Linear regression with L1 regularization is also commonly called Lasso regression.