MAP + prior = Regularization

Έστω ένα σύνολο δεδομένων $(x_1, y_1), \ldots, (x_N, y_N)$ τέτοια, ώστε

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \quad \text{όπου}$$

ε_i: Γκαουσσιανός θόρυβος με $\varepsilon_i \sim N(0, \sigma^2)$.

Βάσει αυτού ισχύει $y_i \sim N(\hat{y}_i, \sigma^2)$, όπου $\hat{y}_i = w_0 + w_1 x_i$

Η βέλτιστη επιλογή των w_0, w_1 προκύπτει μέσω της μεγιστοποίησης της πιθανοφάνειας $\mathcal{L}(w_0, w_1 | y_1, \ldots, y_N)$, η οποία δίνεται ως

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \hat{y}_i)^2\right] \equiv p(y_1, \ldots, y_N | w_0, w_1)$$

Από το Θ. Bayes ισχύει $p(w_0, w_1 | y_1, \ldots, y_N) \sim \mathcal{L} \cdot p(w_0, w_1)$,

όπου $p(w_0, w_1)$: prior για τις παραμέτρους w_0, w_1.
Μεγιστοποιώντας την posterior $p(w_0, w_1 | y_1, \ldots, y_N)$ δεδομένης μιας prior κατανομής για τις παραμέτρους w_0, w_1 προκύπτουν οι γνωστοί regularizers:

- Έστω Gaussian prior, δηλαδή $p(w_i) \sim N(0, s^2)$. Τότε

$$p(\theta | D) \sim p(D | \theta) \cdot p(\theta) \Rightarrow$$

$$\ln p(\theta | D) \sim \ln\left[\prod_{i=1}^{N} N(y_i | w_0, w_1, \sigma^2)\, N(w_0 | 0, s^2)\, N(w_1 | 0, s^2)\right]$$

$$= -\frac{1}{2}\ln(2\pi s^2) - \frac{1}{2}\frac{w_0^2}{s^2} - \frac{1}{2}\ln(2\pi s^2) - \frac{1}{2}\frac{w_1^2}{s^2} + \sum_{i=1}^{N} \ln N(y_i | w_0, w_1, \sigma^2)$$

$$= -\ln(2\pi s^2) - \frac{1}{2}\frac{w_0^2 + w_1^2}{s^2} - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$= -\ln(2\pi s^2) - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{\sigma^2}{s^2}\|\vec{w}\|^2\right]$$

Προκύπτει, λοιπόν,

$$\vec{w} = \underset{\vec{w}}{\operatorname{argmin}}\, p(\theta|D) = \underset{\vec{w}}{\operatorname{argmin}}\left[\sum_i (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{s^2}\|\vec{w}\|^2\right],$$

δηλαδή L2-regularization με $\lambda = \sigma^2/s^2$.

• Έστω Laplacian prior, δηλαδή $p(w_i) \sim L(0, b)$. Τότε

$$\ln p(\theta|D) = \ln\left[\prod_{i=1}^{N} N(y_i|w_0, w_1, \sigma^2)\, L(w_0|0,b)\, L(w_1|0,b)\right]$$

$$= -\ln(2b) - \frac{|w_0|}{b} - \ln(2b) - \frac{|w_1|}{b} - \sum_{i=1}^{N} \ln N(y_i|w_0, w_1, \sigma^2)$$

$$= -2\ln(2b) - \frac{|w_0| + |w_1|}{b} - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$= -2\ln(2b) - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{2\sigma^2}{b}(|w_0| + |w_1|)\right]$$

Προκύπτει, λοιπόν,

$$\vec{w} = \underset{\vec{w}}{\operatorname{argmin}}\, p(\theta|D) = \underset{\vec{w}}{\operatorname{argmin}}\left[\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \frac{2\sigma^2}{b}\sum_{j=0}^{1}|w_j|\right],$$

δηλαδή L1-regularization με $\lambda = 2\sigma^2/b$.