

Lecture 25: Deep neural networks continued

Professor Ilias Bilonis

Regularization through parameter penalties

to avoid overfitting

Regularization terms in loss functions

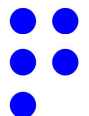
$$J(\theta) = \underbrace{L(\theta)}_{\text{loss func.}} + \underbrace{\lambda R(\theta)}_{\text{regul. term}} + \mu R_1(\theta) + \dots$$

λ \downarrow hyper-parameter
 regularization parameter

could have multiple

$$\begin{aligned}
 + R(\theta) &= \|\theta\|_2^2 = \sum_i \theta_i^2 \\
 + R(\theta) &= \|\theta\|_1 = \sum_i |\theta_i|
 \end{aligned}
 \left. \vphantom{\begin{aligned} + R(\theta) &= \|\theta\|_2^2 \\ + R(\theta) &= \|\theta\|_1 \end{aligned}} \right\} \begin{array}{l} \text{good to} \\ \text{default to} \end{array}$$

;



Bayesian interpretation of regularization

$$\max_{\theta} p(y_{1:n} | x_{1:n}, \theta) \Rightarrow \min_{\theta} L(\theta)$$

$$\text{prior over weights } p(\theta) \Rightarrow \text{posterior } p(\theta | x_{1:n}, y_{1:n}) \propto \text{likelihood } p(y_{1:n} | x_{1:n}, \theta) \text{ prior } p(\theta)$$

$$\text{MAP of } \theta: \max_{\theta} \log p(\theta | x_{1:n}, y_{1:n})$$

$$J(\theta) = -\log p(\theta | x_{1:n}, y_{1:n}) = -\log \frac{p(y_{1:n} | x_{1:n}, \theta) p(\theta)}{L(\theta)} = \underbrace{-\log p(y_{1:n} | x_{1:n}, \theta)}_{L(\theta)} + \underbrace{-\log p(\theta)}_{\lambda R(\theta)}$$

Gaussian prior: $\theta \sim \mathcal{N}(0, \lambda^{-1})$

$$-\log p(\theta) = -\log \mathcal{N}(\theta | 0, \lambda^{-1})$$

$$= -\lambda \underbrace{\|\theta\|_2^2}_{R(\theta)} + \text{const.}, \text{ w.r.t } \theta$$

$$\Rightarrow R(\theta) = \|\theta\|_2^2$$

Gaussian prior
 \rightarrow L2 regularization term