Lecture 25: Deep neural networks continued

Professor Ilias Bilionis

Regularization through parameter penalties

to avoid overfitting

Regularization terms in loss functions

$$J(\theta) = \underbrace{L(\theta)}_{\text{loss funct.}} + \lambda \underbrace{R(\theta)}_{\text{regul. term}} + \mu R_2(\theta) + \dots$$

could have multiple

hyper-parameter

regularizing parameter.

$$+ \quad R(\theta) = \|\theta\|_2^2 = \sum_i \theta_i^2$$

$$+ \quad R(\theta) = \|\theta\|_1 = \sum_i |\theta_i|$$

$$\vdots$$

good to default to

Bayesian interpretation of regularization

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

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

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

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

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

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

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

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

Gaussian prior

↳ L2 regularization term

PREDICTIVE SCIENCE LABORATORY