Problem Set 5 - Regularization

DS542 DL4DS

Spring, 2025
Sicheng Yi (Tiger Yi)
tigeryi@bu.edu

Note: Refer to the equations in the *Understanding Deep Learning* textbook to solve the following problems.

AI/Correction Statement (1 point)

You may use ChatGPT/Generative AI as a resource to help you complete the assignment. However, it must be used constructively to help you understand things you are unsure of, and be built upon with original work.

You must cite your interaction by describing your prompt and the corresponding response. In addition, you must explain all output from the AI that you implement in your assignment. Failure to do so could result in credit deduction.

The official GAIA Policy can be found here.

Moreover, if this is a correction submission after the initial submission, you must provide a reflection on what you learned from the initial submission and how you corrected it.

Answer:

For problem 9.1, I did NOT use any generate AI like the ChatGPT, I typed all the latex equations in the .tex file

For problem 9.5, I did NOT use any generate AI like the ChatGPT, I typed all the latex equations in the .tex file

Problem 9.1 (4 points)

Consider a model where the prior distribution over the parameters is a normal distribution with mean zero and variance σ_ϕ^2 so that

$$Pr(\phi) = \prod_{j=1}^{J} \mathcal{N}(\phi_j|0, \sigma_\phi^2), \tag{1}$$

where j indexes the model parameters. When we apply a prior, we maximize

$$\prod_{i=1}^{I} Pr(y_i|x_i, \phi)Pr(\phi). \tag{2}$$

Show that the associated loss function of this model is equivalent to L2 regularization.

Answer:

$$\phi^{'} = \arg\max_{\phi}[\prod_{i=1}^{I} Pr(y_i|x_i, \phi)Pr(\phi)] \tag{3}$$

$$\phi^{'} = -\log\arg\min_{\phi}[\prod_{i=1}^{I} Pr(y_i|x_i, \phi)Pr(\phi)] \tag{4}$$

$$\phi^{'} = -\arg\min_{\phi}\sum_{i=1}^{I}[\log Pr(y_i|x_i, \phi) + \log Pr(\phi)] \tag{5}$$

$$Pr(\phi) = \prod_{j=1}^{J} \mathcal{N}(\phi_j|0, \sigma_\phi^2) \approx \prod_{j=1}^{J} C \cdot \exp(-\frac{\phi_j^2}{2\sigma_\phi^2}) \tag{6}$$

$$-\log[Pr(\phi)] = \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2 = \lambda\sum_{j=1}^{J}\phi_j^2 \tag{7}$$

$$-\log Pr(y_i|x_i, \phi) = \ell_i[\phi, x_i, y_i] \text{ negative log likelihood} \tag{8}$$

$$\phi^{'} = \arg\min_{\phi}\sum_{i=1}^{I}[\ell_i[\phi, x_i, y_i] + \lambda\sum_{j=1}^{J}\phi_j^2] \tag{9}$$

Therefore, I have shown the loss function is equivalent to the L2 ridge regularization with the last term:

$$\lambda\sum_{j=1}^{J}\phi_j^2$$

Problem 9.5 (4 points)

Show that the weight decay parameter update with decay rate λ:

$$\phi \leftarrow (1-\lambda)\phi - \alpha\frac{\partial L}{\partial \phi}, \tag{10}$$

on the original loss function $L[\phi]$ is equivalent to a standard gradient update using L2 regularization, so that the modified loss function $\tilde{L}[\phi]$ is:

$$\tilde{L}[\phi] = L[\phi] + \frac{\lambda}{2\alpha}\sum_k \phi_k^2, \tag{11}$$

where ϕ represents the parameters, and α is the learning rate.

Answer:

Take derivative of $\tilde{L}$ with respect to ϕ

$$\frac{\partial \tilde{L}}{\partial \phi} = \frac{\partial L}{\partial \phi} + \frac{\lambda}{\alpha}\phi \tag{12}$$

$$\phi \leftarrow \phi - \alpha\frac{\partial \tilde{L}}{\partial \phi} \tag{13}$$

$$\phi \leftarrow \phi - \alpha(\frac{\partial L}{\partial \phi} + \frac{\lambda}{\alpha}\phi) \tag{14}$$

$$\phi \leftarrow \phi - \alpha\frac{\partial L}{\partial \phi} - \lambda\phi \tag{15}$$

$$\phi \leftarrow (1-\lambda)\phi - \alpha\frac{\partial L}{\partial \phi} \tag{16}$$

Therefore, gradient descent on the modified $\tilde{L}$ with L2 regularization is same as applying weight decay λ on the old loss function L