Problem Set 5 - Regularization

DS542 – DL4DS

Spring, 2025

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.

- I did not use generative AI for this assignment.

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:

The likelihood function for the model, given the parameters, is:

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

The prior distribution over the parameters ϕ is:

$$Pr(\phi) = \prod_{j=1}^{J} \mathcal{N}(\phi_j|0, \sigma_\phi^2) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_\phi^2}} \exp\left(-\frac{\phi_j^2}{2\sigma_\phi^2}\right)$$

The objective is to maximize the posterior, which is the product of the likelihood and the prior. We can take the log-likelihood since maximizing the log of a function is equivalent to maximizing the function itself. The log posterior is:

$$\log\left(\prod_{i=1}^{I} Pr(y_i|x_i, \phi) Pr(\phi)\right)$$

$$\sum_{i=1}^{I} \log Pr(y_i|x_i, \phi) + \sum_{j=1}^{J} \log \mathcal{N}(\phi_j|0, \sigma_\phi^2)$$

Expanding the second sum, which is the log of the normal distribution over each ϕ_j:

$$\log \mathcal{N}(\phi_j|0, \sigma_\phi^2) = -\frac{1}{2}\log(2\pi\sigma_\phi^2) - \frac{\phi_j^2}{2\sigma_\phi^2}$$

$$\sum_{j=1}^{J} \log \mathcal{N}(\phi_j|0, \sigma_\phi^2) = -\frac{J}{2}\log(2\pi\sigma_\phi^2) - \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2$$

The full log posterior is:

$$\sum_{i=1}^{I} \log Pr(y_i|x_i, \phi) - \frac{J}{2}\log(2\pi\sigma_\phi^2) - \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2$$

Maximizing the log posterior is equivalent to minimizing the negative log posterior. The term involving $\sum_{j=1}^{J}\phi_j^2$ leads to a penalty on the size of the parameters ϕ_j. So the regularization term is:

$$\frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2$$

This term is the L2 regularization term, where $\frac{1}{2\sigma_\phi^2}$ acts as a regularization parameter.

Therefore, the associated loss function that corresponds to maximizing the posterior includes an L2 regularization term on the parameters ϕ, with the strength of the regularization determined by σ_ϕ^2.

Thus, the loss function can be written as:

$$-\sum_{i=1}^{I} \log Pr(y_i|x_i, \phi) + \frac{1}{2\sigma_\phi^2} \sum_{j=1}^{J} \phi_j^2$$

Problem 9.5 (4 points)

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

$$\phi \leftarrow (1-\lambda)\phi - \alpha \frac{\partial L}{\partial \phi}, \tag{3}$$

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:

$$\tilde{L}[\phi] = L[\phi] + \frac{\lambda}{2\alpha} \sum_k \phi_k^2, \tag{4}$$

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

Answer:

The modified loss function with L2 regularization is:

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

This is the original loss function $L[\phi]$ plus a regularization term $\frac{\lambda}{2\alpha} \sum_k \phi_k^2$, where ϕ_k^2 is the squared value of the parameter ϕ_k and λ is the regularization strength.

To understand how this modified loss $\tilde{L}[\phi]$ leads to the same update rule, we first compute the gradient of the regularized loss function $\tilde{L}[\phi]$:

$$\frac{\partial \tilde{L}[\phi]}{\partial \phi_k} = \frac{\partial L[\phi]}{\partial \phi_k} + \frac{\lambda}{\alpha}\phi_k.$$

The gradient of the loss $L[\phi]$ with respect to ϕ_k remains the same as in the original loss function, while the second term $\frac{\lambda}{\alpha}\phi_k$ is the gradient of the regularization term $\frac{\lambda}{2\alpha} \sum_k \phi_k^2$ with respect to ϕ_k.

The update rule based on this gradient is:

$$\phi_k \leftarrow \phi_k - \alpha \frac{\partial \tilde{L}[\phi]}{\partial \phi_k}.$$

Substituting the gradient of $\tilde{L}[\phi]$ into this update:

$$\phi_k \leftarrow \phi_k - \alpha \left(\frac{\partial L[\phi]}{\partial \phi_k} + \frac{\lambda}{\alpha} \phi_k \right).$$

Simplifying:

$$\phi_k \leftarrow \phi_k - \alpha \frac{\partial L[\phi]}{\partial \phi_k} - \lambda \phi_k.$$

This is equivalent to the original weight decay update rule:

$$\phi_k \leftarrow (1 - \lambda)\phi_k - \alpha \frac{\partial L}{\partial \phi_k}.$$