Chapter 9: Regularisation

Problem 9.1

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

$$Pr(\phi) = \prod_{j=1}^{J} \text{Norm}_{\phi_j}[0, \sigma_\phi^2]$$

where j indexes the model parameters. We now maximise $\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi})Pr(\boldsymbol{\phi})$. Show that the associated loss function of this model is equivalent to L2 regularisation.

The probability density function of a normal distribution $\text{Norm}_{\phi_j}[0, \sigma_\phi^2]$ is given by:

$$Pr(\phi_j) = \frac{1}{\sqrt{2\pi\sigma_\phi^2}} \exp\left(-\frac{\phi_j^2}{2\sigma_\phi^2}\right)$$

Therefore, the joint prior over all parameters ϕ is:

$$\begin{aligned} Pr(\phi) &= \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_\phi^2}} \exp\left(-\frac{\phi_j^2}{2\sigma_\phi^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\right)^J \exp\left(-\sum_{j=1}^{J} \frac{\phi_j^2}{2\sigma_\phi^2}\right) \end{aligned}$$

Taking the logarithm of the joint prior, we have:

$$\begin{aligned} \log Pr(\phi) &= -\frac{J}{2}\log(2\pi\sigma_\phi^2) - \sum_{j=1}^{J} \frac{\phi_j^2}{2\sigma_\phi^2} \\ &= -\frac{J}{2}\log(2\pi\sigma_\phi^2) - \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J} \phi_j^2 \end{aligned}$$

Maximising the joint prior is equivalent to minimising the negative of the joint prior. Therefore, the loss function is:

$$L[\phi] = \frac{J}{2}\log(2\pi\sigma_\phi^2) + \frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J} \phi_j^2$$

Where the penalty term is $\frac{1}{2\sigma_\phi^2}\sum_{j=1}^{J}\phi_j^2$, which is equivalent to L2 regularisation. ($\lambda = \frac{1}{2\sigma_\phi^2}$)

Problem 9.2

How do the gradients of the loss function change when L2 regularisation (equation 9.5) is added?

The loss function with L2 regularisation is given by:

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

Meaning that the gradient update rule now becomes:

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

where α is the learning rate, and λ is the regularisation parameter. Using L2 regularisation therefore encourages the learning algorithm to prefer smaller parameter values, reducing overfitting and improving model generalization.

Problem 9.3

Consider a linear regression model $y = \phi_0 + \phi_1 x$ with input x, output y, and parameters ϕ_0 and ϕ_1. Assume we have I training examples $\{x_i, y_i\}$ and use a least squares loss. Consider adding Gaussian noise with mean zero and variance σ_x^2 to the inputs x_i at each training iteration. What is the expected gradient update?

Adding Gaussian noise to the inputs x_i at each training iteration means that the loss function becomes:

$$\begin{aligned}
\tilde{L} &= \sum_{i=1}^{I}(\phi_0 + \phi_1(x_i + \epsilon_i) - y_i)^2 \\
&= \sum_{i=1}^{I}(\phi_0 + \phi_1 x_i - y_i + \phi_1\epsilon_i)^2 \\
&= \sum_{i=1}^{I}(\phi_0 + \phi_1 x_i - y_i)^2 + \phi_1^2\sum_{i=1}^{I}\epsilon_i^2 + \sum_{i=1}^{I}(\phi_0 + \phi_1 x_i - y_i)\epsilon_i
\end{aligned}$$

Taking expectations, and noting that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma_x^2$, the expected gradient update is:

$$\begin{aligned}\mathbb{E}[\tilde{L}] &= \sum_{i=1}^{I}(\phi_0 + \phi_1 x_i - y_i)^2 + \phi_1^2 \sum_{i=1}^{I}\sigma_x^2 \\ &= L + (I\sigma_x^2)\phi_1^2\end{aligned}$$

which is equivalent to applying L2 regularisation with constant $\lambda = I\sigma_x^2$.

Problem 9.4

Derive the loss function for multiclass classification when we use label smoothing so that the target probability distribution has 0.9 at the correct class and the remaining probability mass of 0.1 is divided between the remaining $D_o - 1$ classes.

The probability of the data is now:

$$Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi}) = 0.9 \cdot \text{softmax}_{y_i}[\mathbf{f}(\mathbf{x}_i, \boldsymbol{\phi})] + \sum_{z \in \{1 \ldots D_o\} \setminus y_i}^{max} \frac{0.1}{D_o - 1} \cdot \text{softmax}_z[\mathbf{f}(\mathbf{x}_i, \boldsymbol{\phi})]$$

and the loss function is the negative log probability of this quantity.

Problem 9.5

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

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

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

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

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

To perform gradient descent on the modified loss function $\tilde{L}[\phi]$, we need to compute the gradient of $\tilde{L}[\phi]$ with respect to ϕ:

$$\frac{\partial \tilde{L}}{\partial \phi} = \frac{\partial L}{\partial \phi} + \frac{\lambda}{2\alpha} \cdot 2\phi = \frac{\partial L}{\partial \phi} + \frac{\lambda}{\alpha}\phi$$

Therefore, the gradient update rule becomes:

$$\begin{aligned}\phi &\leftarrow \phi - \alpha\left(\frac{\partial L}{\partial \phi} + \frac{\lambda}{\alpha}\phi\right) \\ &\leftarrow \phi - \alpha\frac{\partial L}{\partial \phi} - \lambda\phi \\ &\leftarrow (1-\lambda)\phi - \alpha\frac{\partial L}{\partial \phi}\end{aligned}$$

Problem 9.6

Consider a model with parameters $\boldsymbol{\phi} = [\boldsymbol{\phi}_0, \boldsymbol{\phi}_1]^T$. Draw the $L0$, $L\frac{1}{2}$, and $L1$ regularisation terms in a similar form to figure 9.1b. The LP regularisation term is $\sum_{d=1}^{D} |\phi_d|^P$.