2 Supervised Learning

Problems

Problem 2.1 To walk "downhill" on the loss function (equation 2.5), we measure its gradient with
respect to the parameters
ϕ_0 and ϕ_1. Calculate expressions for the slopes $\frac{\partial L}{\partial \phi_0}$ and $\frac{\partial L}{\partial \phi_1}$.

Given that

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

$$\frac{\partial L}{\partial \phi_0} = 2 \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)$$

$$\frac{\partial L}{\partial \phi_1} = 2 \sum_{i=1}^{I} x_i (\phi_0 + \phi_1 x_i - y_i)$$

Problem 2.2 Show that we can find the minimum of the loss function in closed form by setting the expression for the derivatives from problem 2.1 to zero and solving for ϕ_0 and ϕ_1. Note that this works for linear regression but not for more complex models; this is why we use iterative model fitting methods like gradient descent (figure 2.4).

Since the original loss function is a quadratic function of ϕ_0 and ϕ_1, its surface is concave and has a single global minimum. We can find the global minimum by solving the partial gradient w.r.t. ϕ_0 and ϕ_1 to 0:

$$\frac{\partial L}{\partial \phi_0} = 2\sum_{i=1}^{I}(\phi_0 + \phi_1 x_i - y_i) = 0$$
$$\frac{\partial L}{\partial \phi_1} = 2\sum_{i=1}^{I}x_i(\phi_0 + \phi_1 x_i - y_i) = 0$$

Restructuring the equation:

$$I\phi_0 + \sum_i x_i\phi_1 = \sum_i y_i$$
$$\sum_i x_i\phi_0 + \sum_i x_i^2\phi_1 = \sum_i x_i y_i$$

The solution for this system of equations is:

$$\phi_0 = \frac{1}{I\sum_i x_i^2 - (\sum_i x_i)^2}\left[(\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i)\right]$$
$$\phi_1 = \frac{1}{I\sum_i x_i^2 - (\sum_i x_i)^2}\left[-(\sum_i x_i)(\sum_i y_i) + I\sum_i x_i y_i\right]$$

Problem 2.3 Consider reformulating linear regression as a generative model, so we have $x = g[y, \boldsymbol{\phi}] = \phi_0 + \phi_1 y$. What is the new loss function? Find an expression for the inverse function $y = g^{-1}[x, \boldsymbol{\phi}]$ that we would use to perform inference. Will this model make the same predictions as the discriminative version for a given training dataset $\{x_i, y_i\}$? One way to establish this is to write code that fits a line to three data points using both methods and see if the result is the same.

The new loss function is the sum of square losses between the data points $\{x_i, y_i\}$ and the generated points $\{g[y_i, \boldsymbol{\phi}], y_i\}$.

$$L[\boldsymbol{\phi}] = \sum_i (x_i - g[y_i, \boldsymbol{\phi}])^2$$

The inverse function can be computed by solving the equation $x = \phi_0 + \phi_1 y$ w.r.t. y:

$$y = g^{-1}[x, \boldsymbol{\phi}] = \frac{x - \phi_0}{\phi_1}$$

We can think of an extreme case to show that discriminative and generative model **does not** necessarily produce the same result. If the data points are aligned is a horizontal line in (x, y) plane, the discriminative model can predict a perfect result with $\phi_1 = 0$. However generative model cannot do so, since the model function g can only produce a single value of x for a given y so it can only fit at most single point among the data points.