Exercise Sheet 2

Neural Computation
University of Birmingham

Week 2 2025

Exercise 1. Consider the the problem of finding an accurate temperature estimate $\hat{x}$ from a series of noisy measurements $\mathbf{x} = (x_1, \ldots, x_n)$. We want to find the estimate that minimises the sum of squared errors

$$x^* = \arg\min_{\hat{x}} L(\hat{x}, \mathbf{x}), \tag{1}$$

where

$$L(\hat{x}, \mathbf{x}) = \sum_{i=1}^{n} (\hat{x} - x_i)^2. \tag{2}$$

For this particular problem, there is an analytical solution for x^*, that is, a solution that we can find without requiring optimisation methods like gradient descent. Find this solution. Here are some helpful tips:

- Find the minimum of $L(\hat{x}, \mathbf{x})$ by setting its derivative $\frac{\partial}{\partial \hat{x}} L(\hat{x}, \mathbf{x}) = 0$ and solving for $\hat{x}$.
- The derivative of the sum is the sum of the derivatives of the summands.
- Remember that for a quadratic equation $(a - b)^2 = a^2 - 2ab + b^2$.

Solution:

We are asked to find the estimate $\hat{x}$ that minimises the loss function

$$L(\hat{x}, \mathbf{x}) = \sum_{i=1}^{n} (\hat{x} - x_i)^2.$$

To find the minimum, we take the derivative of $L(\hat{x}, \mathbf{x})$ with respect to $\hat{x}$ and set it to zero:

$$\begin{aligned}
\frac{\partial}{\partial \hat{x}} L(\hat{x}, \mathbf{x}) &= \frac{\partial}{\partial \hat{x}} \sum_{i=1}^{n} (\hat{x} - x_i)^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \hat{x}} (\hat{x} - x_i)^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \hat{x}} \left(\hat{x}^2 - 2\hat{x}x_i + x_i^2 \right) \\
&= \sum_{i=1}^{n} 2\hat{x} - 2x_i \\
&= \sum_{i=1}^{n} 2\hat{x} - \sum_{i=1}^{n} 2x_i \\
&= 2n\hat{x} - \sum_{i=1}^{n} 2x_i.
\end{aligned}$$

Setting the derivative equal to zero gives

$$2n\hat{x} - \sum_{i=1}^{n} 2x_i = 0.$$

Simplifying,

$$n\hat{x} = \sum_{i=1}^{n} x_i.$$

Solving for $\hat{x}$,

$$\hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Therefore, the analytical solution is

$$x^* = \hat{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

which is simply the **sample mean** of the measurements.

Exercise 2. Scientists interested in the penguin population on an Antarctic island counted the number of animals three times, resulting in population numbers of 332, 353, and 419, obtained in the years 2010, 2012, and 2016, respectively.

They want to fit a linear model (a perceptron without the sign function) to the data, describing the development of penguin numbers over time. To make the calculation easier, they express the years relative to the year of the first measurement. They initialise their model parameters by hand using a rough guess with slope $w^0 = 10$ and intercept $b^0 = 350$.

(a) According to this rough initial model, what would be the projected number of penguins in the year 2025?

(b) The scientists want to update their model using gradient descent, trying to minimise the sum of absolute errors

$$\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

as their loss function. Compute the loss of this model with the initial parameters w^0, b^0.

(c) Compute the first gradient update for w and b, using a learning rate of $\eta = 0.5$.

Tip: The derivative of the absolute value function is not defined at zero, but for non-zero values it can be written as:

$$\frac{d}{dx}|x| = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Use this to compute the gradient of the loss with respect to w and b.

(d) Compute the loss after the gradient step.

Solution:

Here are the datapoints we are considering

$$(0, 332), \quad (2, 353), \quad (6, 419)$$

where the first coordinate is the number of years since 2010.

The initial model is:

$$\hat{y} = w^0 x + b^0 = 10x + 350$$

(a) In the year 2025, the number of years since 2010 is $x = 15$. Using the model:

$$\hat{y} = 10 \cdot 15 + 350 = 150 + 350 = 500$$

So, the projected number of penguins in 2025 is **500**.

(b) Compute the predicted values for each data point:

$$\hat{y}_1 = 10 \cdot 0 + 350 = 350$$
$$\hat{y}_2 = 10 \cdot 2 + 350 = 370$$
$$\hat{y}_3 = 10 \cdot 6 + 350 = 410$$

Compute the absolute errors:

$$|\hat{y}_1 - y_1| = |350 - 332| = 18$$
$$|\hat{y}_2 - y_2| = |370 - 353| = 17$$
$$|\hat{y}_3 - y_3| = |410 - 419| = 9$$

Total loss:

$$L = 18 + 17 + 9 = \boxed{44}$$

(c) Compute the gradients of the loss with respect to w and b. The derivatives of the loss can be written as

$$\frac{\partial}{\partial w} \sum_{i=1}^{n} |\hat{y}_i - y_i| = \sum_{i=1}^{n} \frac{\partial}{\partial w} |\hat{y}_i - y_i| = \sum_{i=1}^{n} \frac{\partial}{\partial w} |wx_i + b - y_i|$$

and

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} |\hat{y}_i - y_i| = \sum_{i=1}^{n} \frac{\partial}{\partial b} |\hat{y}_i - y_i| = \sum_{i=1}^{n} \frac{\partial}{\partial b} |wx_i + b - y_i|$$

For each data point (x_i, y_i), the gradient of the absolute error is:

$$\frac{\partial}{\partial w} |\hat{y}_i - y_i| = \text{sign}(\hat{y}_i - y_i) \cdot x_i, \quad \frac{\partial}{\partial b} |\hat{y}_i - y_i| = \text{sign}(\hat{y}_i - y_i)$$

Using the signs:

$$\text{sign}(\hat{y}_1 - y_1) = \text{sign}(350 - 332) = +1$$
$$\text{sign}(\hat{y}_1 - y_1) = \text{sign}(370 - 353) = +1$$
$$\text{sign}(\hat{y}_1 - y_1) = \text{sign}(410 - 419) = -1$$

Compute gradients:

$$\frac{\partial L}{\partial w} = 1 \cdot 0 + 1 \cdot 2 + (-1) \cdot 6 = -4$$
$$\frac{\partial L}{\partial b} = 1 + 1 - 1 = 1$$

Apply gradient descent update with learning rate $\eta = 0.5$:

$$w^1 = w^0 - \eta \cdot \frac{\partial L}{\partial w} = 10 - 0.5 \cdot (-4) = 10 + 2 = \boxed{12}$$
$$b^1 = b^0 - \eta \cdot \frac{\partial L}{\partial b} = 350 - 0.5 \cdot 1 = 350 - 0.5 = \boxed{349.5}$$

(d) Using the updated model $\hat{y} = 12x + 349.5$, compute new predictions:

$$\hat{y}_1 = 12 \cdot 0 + 349.5 = 349.5$$
$$\hat{y}_2 = 12 \cdot 2 + 349.5 = 24 + 349.5 = 373.5$$
$$\hat{y}_3 = 12 \cdot 6 + 349.5 = 72 + 349.5 = 421.5$$

Compute new absolute errors:

$$|\hat{y}_1 - y_1| = |349.5 - 332| = 17.5$$
$$|\hat{y}_2 - y_2| = |373.5 - 353| = 20.5$$
$$|\hat{y}_3 - y_3| = |421.5 - 419| = 2.5$$

New total loss:

$$L = 17.5 + 20.5 + 2.5 = \boxed{40.5}$$

Exercise 3. There are two species of penguins living on the island: Adelie and Gentoo penguins. Consider a perceptron model for classifying the penguin species based on flipper length in cm (x_1) and body weight in kg (x_2). The model uses a logistic function as the activation function and outputs the probability that a penguin is a Gentoo penguin. Its parameters are $\mathbf{w} = (1, 1)$ and $b = -25$.

(a) Consider three penguins with flipper lengths of 22, 22, and 23 cm, and body masses of 2, 3, and 4 kg, respectively. Compute their probabilities of being Gentoo penguins according to the model.

(b) We want to change the weight vector to be $\mathbf{w}' = (2, 2)$ but adjust $b' = -50$. Compute the probabilities for the three penguins again using the new values. Does the new model have a different decision boundary? How can you describe the influence of the length of the weight vector on the model?

Solution:

We are given two logistic regression models:
- **Model 1**: $\mathbf{w} = (1, 1), \quad b = -25$ - **Model 2**: $\mathbf{w}' = (2, 2), \quad b' = -50$

The logistic function is:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \text{where } z = w_1 x_1 + w_2 x_2 + b$$

(a) For the three penguins:

- Penguin 1: $x_1 = 22$, $x_2 = 2$

$$\text{Model 1: } z = 22 + 2 - 25 = -1 \Rightarrow \qquad P(\text{Gentoo}) \approx \boxed{0.2689}$$
$$\text{Model 2: } z = 2 \cdot 22 + 2 \cdot 2 - 50 = -2 \Rightarrow \qquad P(\text{Gentoo}) \approx \boxed{0.1192}$$

- Penguin 2: $x_1 = 22$, $x_2 = 3$

$$\text{Model 1: } z = 22 + 3 - 25 = 0 \Rightarrow \qquad P(\text{Gentoo}) = \boxed{0.5}$$
$$\text{Model 2: } z = 2 \cdot 22 + 2 \cdot 3 - 50 = 0 \Rightarrow \qquad P(\text{Gentoo}) = \boxed{0.5}$$

- Penguin 3: $x_1 = 23$, $x_2 = 4$

$$\text{Model 1: } z = 23 + 4 - 25 = 2 \Rightarrow \qquad P(\text{Gentoo}) \approx \boxed{0.8808}$$
$$\text{Model 2: } z = 2 \cdot 23 + 2 \cdot 4 - 50 = 4 \Rightarrow \qquad P(\text{Gentoo}) \approx \boxed{0.9820}$$

(b) The decision boundary for both models is defined by:

$$\text{Model 1: } 1x_1 + 1x_2 = 25$$

and

$$\text{Model 2: } 2x_1 + 2x_2 = 50 \Rightarrow x_1 + x_2 = 25$$

So, both models have the **same decision boundary**. If we scale the weight vector, and scale the bias by the same factor, the decision boundary does not change.

Influence of weight vector length: However, increasing the length of the weight vector makes the sigmoid function **steeper**, resulting in: - More confident predictions (closer to 0 or 1) for inputs near the decision boundary. - Less gradual transitions between classes.

Although both models classify the same boundary, **Model 2 produces more extreme probabilities** due to the larger weight magnitudes. This shows that the **length of the weight vector controls the sharpness of the classification**, while the **bias determines the position of the decision boundary**.