Assignment 1

November 25, 2025

1 Evaluate θ^1

Model and Loss

The model is given by:

$$h(x_1, x_2) = \sigma(z), \quad \text{where } z = b + w_1 x_1 + w_2 x_2$$

The Sigmoid function is:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The Mean Squared Error (MSE) loss is:

$$L = \frac{1}{2}(y - h)^2 = \frac{1}{2}(h - y)^2$$

Gradient Calculation

The gradients are:

$$\frac{\partial L}{\partial b} = (h - y)\sigma'(z), \quad \frac{\partial L}{\partial w_i} = (h - y)\sigma'(z)\, x_i$$

where the derivative of the sigmoid function is:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$

Numerical Evaluation

Given the data point $(x_1, x_2, y) = (1, 2, 3)$ and the initial parameters $\theta^0 = (b, w_1, w_2) = (4, 5, 6)$.

First, calculate z:

$$z = b + w_1 x_1 + w_2 x_2 = 4 + 5(1) + 6(2) = 4 + 5 + 12 = 21$$

The update rule for Gradient Descent is $\theta^1 = \theta^0 - \alpha \nabla_\theta L$.

Let $G = (\sigma(z) - y)\sigma'(z) = (\sigma(21) - 3)\sigma(21)(1 - \sigma(21))$.

The updated parameters $\theta^1 = (b^1, w_1^1, w_2^1)$ are:

$$\begin{aligned}
b^1 &= b^0 - \alpha \frac{\partial L}{\partial b} = 4 - \alpha\big(\sigma(21) - 3\big)\,\sigma(21)\big(1 - \sigma(21)\big) \cdot 1 \\
w_1^1 &= w_1^0 - \alpha \frac{\partial L}{\partial w_1} = 5 - \alpha\big(\sigma(21) - 3\big)\,\sigma(21)\big(1 - \sigma(21)\big) \cdot 1 \\
w_2^1 &= w_2^0 - \alpha \frac{\partial L}{\partial w_2} = 6 - \alpha\big(\sigma(21) - 3\big)\,\sigma(21)\big(1 - \sigma(21)\big) \cdot 2
\end{aligned}$$

---

2 Properties of the Sigmoid Function

2.(a) Finding the Expressions

Given $\sigma(x) = \frac{1}{1+e^{-x}}$.

First Derivative $\sigma'(x)$

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x))$$

Second Derivative $\sigma''(x)$

$$\begin{aligned}\sigma''(x) &= \frac{d}{dx}[\sigma(x)(1-\sigma(x))] \\ &= \sigma'(x)(1-\sigma(x)) + \sigma(x)(-\sigma'(x)) \\ &= \sigma'(x)(1-2\sigma(x)) \\ &= \sigma(x)(1-\sigma(x))(1-2\sigma(x))\end{aligned}$$

Third Derivative $\sigma^{(3)}(x)$

We use the expression $\sigma''(x) = 2\sigma^3(x) - 3\sigma^2(x) + \sigma(x)$:

$$\begin{aligned}\sigma^{(3)}(x) &= \frac{d}{dx}[2\sigma^3 - 3\sigma^2 + \sigma] \\ &= 2(3\sigma^2\sigma') - 3(2\sigma\sigma') + \sigma' \\ &= \sigma'(6\sigma^2 - 6\sigma + 1) \\ &= \sigma(x)(1-\sigma(x))(6\sigma^2(x) - 6\sigma(x) + 1)\end{aligned}$$

2.(b) Relation between Sigmoid and Hyperbolic Functions

The hyperbolic functions are:

$$\begin{aligned}\sinh(x) &= \frac{e^x - e^{-x}}{2} \\ \cosh(x) &= \frac{e^x + e^{-x}}{2} \\ \tanh(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}}\end{aligned}$$

From the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, we can derive e^x and e^{-x} in terms of σ:

$$\sigma(x) = \frac{e^x}{e^x+1} \implies e^x\sigma(x) + \sigma(x) = e^x \implies \sigma(x) = e^x(1-\sigma(x)) \implies e^x = \frac{\sigma(x)}{1-\sigma(x)}$$

And $e^{-x} = \frac{1}{e^x} = \frac{1-\sigma(x)}{\sigma(x)}$. For brevity, let $\sigma = \sigma(x)$.

(1) $\sinh(x)$ **in terms of** $\sigma(x)$

$$\begin{aligned}\sinh(x) &= \frac{e^x - e^{-x}}{2} = \frac{1}{2}\left(\frac{\sigma}{1-\sigma} - \frac{1-\sigma}{\sigma}\right) \\ &= \frac{1}{2}\frac{\sigma^2 - (1-\sigma)^2}{\sigma(1-\sigma)} = \frac{1}{2}\frac{\sigma^2 - (1-2\sigma+\sigma^2)}{\sigma(1-\sigma)} \\ &= \frac{1}{2}\frac{2\sigma - 1}{\sigma(1-\sigma)} = \frac{2\sigma(x) - 1}{2\sigma(x)(1-\sigma(x))}\end{aligned}$$

(2) $\cosh(x)$ **in terms of** $\sigma(x)$

$$\begin{aligned}\cosh(x) &= \frac{e^x + e^{-x}}{2} = \frac{1}{2}\left(\frac{\sigma}{1-\sigma} + \frac{1-\sigma}{\sigma}\right) \\ &= \frac{1}{2}\frac{\sigma^2 + (1-\sigma)^2}{\sigma(1-\sigma)} = \frac{1}{2}\frac{\sigma^2 + 1 - 2\sigma + \sigma^2}{\sigma(1-\sigma)} \\ &= \frac{2\sigma^2(x) - 2\sigma(x) + 1}{2\sigma(x)(1-\sigma(x))}\end{aligned}$$

(3) $\tanh(x)$ **in terms of** $\sigma(x)$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

We can also use the identity relating $\tanh(x)$ and $\sigma(2x)$:

$$\begin{aligned}
\sigma(2x) &= \frac{1}{1+e^{-2x}} = \frac{e^{2x}}{e^{2x}+1} \\
2\sigma(2x) - 1 &= \frac{2e^{2x}}{e^{2x}+1} - \frac{e^{2x}+1}{e^{2x}+1} = \frac{e^{2x}-1}{e^{2x}+1} \\
&= \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x)
\end{aligned}$$

The relations are:

$$\sigma(x) = \frac{1}{2}\left(1 + \tanh\left(\frac{x}{2}\right)\right), \qquad \tanh(x) = 2\sigma(2x) - 1$$

---

3 Some Questions

Q1. Why do we usually begin with a linear model?

Key Advantages of Linear Models

(**1**) **Simplicity and Interpretability**: Linear models are mathematically simple. Their parameters (coefficients) directly indicate the strength and direction of the relationship between predictors and the target variable, making them highly **interpretable**.

(**2**) **Computational Efficiency**: They are fast to train and require fewer computational resources, especially compared to complex non-linear models.

(**3**) **Well-Understood Statistical Properties**: Linear models are the foundation of many statistical theories. Their properties, such as variance and bias, are well-understood, allowing for robust statistical inference.

(**4**) **Foundation for More Complex Models**: They serve as the core component of many advanced models (e.g., the input layer of a neural network is a linear combination of features).

When to Use a Linear Model

- When the relationship is likely **linear**.
- For establishing **baseline performance** before moving to more complex models.
- When **interpretability** and ease of communication are priorities.

Limitations

The primary limitation is the assumption of a **linear relationship**. If the underlying data is complex and non-linear, a linear model will likely **underfit** and fail to capture the true patterns.

Q2. What is the difference between BGD, SGD, and Mini-Batch Gradient Descent, and what are their advantages?

These are three main variants of the Gradient Descent optimization algorithm, differing in the amount of data used to calculate the gradient in each update step.

(**1**) **Batch Gradient Descent (BGD)**

- **Process**: Uses the **entire dataset** to compute the gradient of the loss function.
- **Advantages**:
  - Stable convergence because the gradient is the true average across all data.
  - Predictable updates lead to a smooth convergence path toward the minimum.
- **Disadvantages**: Very slow on large datasets; each step is computationally expensive.
- **Use Case**: Small to moderate-sized datasets.

(**2**) **Stochastic Gradient Descent (SGD)**

- **Process**: Uses **only one random example** (or a single data point) at a time to calculate the gradient.
- **Advantages**:
  - Much faster per update, suitable for **very large or streaming datasets**.
  - The noise in the updates can help the model **escape shallow local minima**.
- **Disadvantages**: Updates are noisy and highly fluctuating; convergence is less stable and often requires learning rate decay.
- **Use Case**: Very large/streaming datasets where speed is crucial.

(**3**) **Mini-Batch Gradient Descent (MBGD)**

- **Process**: Uses a **small, random subset** of the data (the mini-batch) to compute the gradient. This is the most common approach in deep learning.
- **Advantages**:
  - **Balance**: Strikes a balance between the speed of SGD and the stability of BGD.
  - Efficiency: Can **exploit parallel processing** (e.g., on GPUs/TPUs) with optimized matrix operations.
  - Stability: Less noisy than SGD, leading to more reliable convergence.
- **Disadvantages**: Requires tuning the optimal **batch size**.
- **Use Case**: Most deep learning tasks.