THỰC HÀNH 2: CÁC PHƯƠNG PHÁP PHÂN LOẠI TUYẾN TÍNH

Bài 1: Sử dụng Numpy xây dựng phương pháp Logistic Regression và Gradient Descent. Tiến hành huấn luyện phương pháp Logistic Regression và trực quan hoá giá trị của hàm loss trong toàn bộ quá trình huấn luyện.

Bộ dữ liệu sử dụng: Predict students' dropout and academic success (link: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success). Ở bài tập này, các bạn lưu ý thực hiện bài toán phân loại 2 lớp với lớp graduate và non-graduate (tức là xem nhãn dropout và enroll là non-graduate).

Gợi ý:

- Xây dựng hàm xác định giá trị hàm mất mát

$$L(\theta, y, X) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right)log\left(1 - \hat{y}^{(i)}\right)\right]$$

trong đó $\hat{y} = \sigma(X\theta^T)$.

- Xây dựng thuật toán Gradient Descent

```
θ := θ0                    // Khởi tạo trọng số
Repeat {
    θ := θ - lr*dL(θ,y,X)/dθ
}
```

Trong đó đạo hàm riêng của $L(\theta, y, X)$ theo θ được xác định bởi:

$$\frac{dL(\theta, y, X)}{d\theta} = \frac{1}{m}X^T(\hat{y} - y)$$

lr là *learning rate*.

Bài 2: Sử dụng Numpy xây dựng phương pháp Softmax Regression và Gradient Descent. Tiến hành huấn luyện phương pháp Logistic Regression và trực quan hoá giá trị của hàm loss trong toàn bộ quá trình huấn luyện.

Bộ dữ liệu sử dụng: Predict students' dropout and academic success. Ở bài tập này, các bạn lưu ý thực hiện bài toán phân loại 3 lớp với lớp graduate, dropout và enroll.

Gợi ý:

- Xây dựng hàm mất mát

$$L(\theta, y, X) = \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{\boldsymbol{C}} y_{i,c} log s_{i,c}$$

với $y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,\boldsymbol{C}})$ là one-hot vector, tức là $y_{i,c} = 1$ và $y_{i,j} = 0\ \forall j \neq c$, $\boldsymbol{C}$ là tổng số lớp cần phân loại.

- Xây dựng phương pháp Gradient Descent tương tự như **Bài 2**. Tuy nhiên đạo hàm riêng của $L(\theta, y_c, X)$ theo θ được xác định bởi:

$$\frac{dL(\theta, y_c, X)}{d\theta} = \frac{dL(\theta, y_c, X)}{dz}\frac{dz}{d\theta} = \frac{1}{m}X^T(s - y)$$

trong đó y là one-hot vector, $s = softmax(\theta^T X)$.

Bài 3: Sử dụng các thư viện Machine Learning (Sklearn hoặc Skorch) thực thi lại 02 phương pháp Logistic Regression và Softmax Regression.