

## THỰC HÀNH 3: CÁC PHƯƠNG PHÁP HỒI QUY VÀ PHÂN LOẠI TUYẾN TÍNH

### Hướng dẫn nộp bài:

Nội dung nộp sẽ bao gồm:

- 01 file báo cáo.
- 01 file zip chứa toàn bộ source code.

Đặt tên là **MSSV\_BaiThucHanh3.zip** và nộp lên course.

**Bài 1 (3 điểm):** Sử dụng Numpy xây dựng phương pháp hồi quy tuyến tính và tiến hành huấn luyện trên bộ dữ liệu California House Price (link: <https://www.kaggle.com/datasets/camnugent/california-housing-prices>). Lưu ý bỏ qua thuộc tính Ocean Proximity và giá trị cần hồi quy là Median House Value.

Gợi ý: tham số tối ưu  $\hat{\theta}$  được xác định bởi  $\hat{\theta} = (X^T X)^{-1} X^T y$ .

**Bài 2 (3 điểm):** 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](https://archive.ics.uci.edu/dataset/697/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

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

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

- Xây dựng thuật toán Gradient Descent

```
W :=  $\theta_0$            // Khởi tạo trọng số
Repeat {
```

$$\theta := \theta - lr * \frac{dL(\theta, y, X)}{d\theta}$$

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

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

$lr$  là *learning rate*.

**Bài 3 (2 điểm):** 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  $L(\theta, y, X) = \frac{1}{m} \sum_{i=1}^m \sum_{c=1}^C y_{i,c} \log s_{i,c}$  với  $y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,C})$  là one-hot vector, tức là  $y_{i,c} = 1$  và  $y_{i,j} = 0 \forall j \neq c$ ,  $C$  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  $\theta$  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,  $s = \text{softmax}(\theta^T X)$ .

**Bài 4 (2 điểm):** Sử dụng các thư viện Machine Learning (Sklearn hoặc Skorch) thực thi lại 3 phương pháp Linear Regression, Logistic Regression và Softmax Regression.