**Bài 4: HUẤN LUYỆN MÔ HÌNH VÀ LỖI**

*Hướng dẫn 1: Phân chia tập dữ liệu huấn luyện và tập dữ liệu kiểm thử*

Cho tập dữ liệu ban đầu 𝐷 = {(𝑋, 𝑦)|𝑋 ∈ 𝑅𝑚×𝑛 𝑣à 𝑦 ∈ 𝑅𝑚} được chia thành 2, gồm:

Tập dữ liệu huấn luyện 𝐷𝑡𝑟𝑎𝑖𝑛 thông thường bằng 70% *D*;

Tập dữ liệu kiểm thử 𝐷𝑡𝑒𝑠𝑡 thông thường bằng 30% *D*.

Thư viện sklearn cung cấp hàm hỗ trợ việc phân chia này. Cú pháp đầy đủ của hàm train\_test\_split() ở [link](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html).

Ở hình thức đơn giản nhất, việc phân chia *D* theo tỉ lệ 70% - 30% được tiến hành như sau:

D = np.loadtxt(os.path.join(folder, filename), delimiter=',') X, y = D[:,:-1], D[:,-1]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y,

test\_size=0.30, random\_state=15)

Câu hỏi:

1. Hãy liệt kê tất cả tham số của hàm train\_test\_split() và ý nghĩa của chúng.

Trả lời: Các tham số của hàm train\_test\_split():

- **\*arrays**: các *list, numpy arrays, ma trận scipy-sparse* hoặc *dataframes* của pandas với cùng độ dài (len) hoặc shape[0] – Dữ liệu đầu vào mà bạn muốn tách thành tập huấn luyện và tập kiểm thử

- **test\_size**: Kiểu dữ liệu *float* hoặc *int* và mặc định là None. Nếu đối số là float thì nó phải nằm giữa 0.0 và 1.0 và thể hiện tỉ lệ giữa tập kiểm thử và tập dữ liệu ban đầu. Nếu đối số là int, thì nó thể hiện số điểm dữ liệu của tập kiểm thử. Nếu đối số là None (mặc định), thì kích thước tập kiểm thử sẽ được thiết lập theo tham số train\_size. Nếu test\_szie = None và train\_size = None, thì sẽ tự động thiết lập test\_size=0.25

- **train\_size**: Kiểu dữ liệu *float* hoặc *int* và mặc định là None. Nếu đối số là float thì nó phải nằm giữa 0.0 và 1.0 và thể hiện tỉ lệ giữa tập huấn luyện và tập dữ liệu ban đầu. Nếu đối số là int, thì nó thể hiện số điểm dữ liệu của tập huấn luyện. Nếu đối số là None (mặc định), thì kích thước tập huấn luyện sẽ được thiết lập theo tham số test\_size. Nếu test\_szie = None và train\_size = None, thì sẽ tự động thiết lập test\_size=0.25

- **random\_state**: Kiểu dữ liệu *int*, mặc định là None. Xác định cách lấy dữ liệu trước khi tách dữ liệu. Cụ thể, nếu một số nguyên (int) được truyền vào (random\_state =5) việc tách tập dữ liệu ban đầu nhiều lần luôn cho kết quả giống nhau. Nếu mặc định (None), thì mỗi lần tách tập dữ liệu sẽ thu được tập huấn luyện và tập kiểm thử khác nhau.

- **shuffle**: Kiểu dữ liệu *bool*, mặc định True. Xét việc có cần xáo trộn dữ liệu trước khi tách không.

- **stratify**: Kiểu dữ liệu giống *array*, mặc định None. Nếu đối số không phải None thì nó sẽ xác định cách tách dữ liệu giống như cách lấy mẫu phân tầng trong thống kê (Một quần thể có thể được phân chia thành các quần thể con)

2. Tại sao khi thực hiện chia train – test, nên lựa chọn tham số ***shuffle = True***? (lưu ý đây là giá trị mặc định)

Trả lời: Tại vì trong nhiều trường hợp dữ liệu có thể được sắp xếp dựa vào đặc trưng nào đó. Ví dụ khi ta thu thập dữ liệu liên quan đến con người trên thành phố Huế thì việc thu thập dữ liệu sẽ được sắp xếp theo từng phường chẳng hạn. Nếu ta không lấy ngẫu nhiên, thì việc lấy dữ liệu có đặc trưng nào đó gần nhau có thể tương đồng dẫn đến không có tính tổng quát (generalization). Để hạn chế điều đó xảy ra, ta sẽ tiến hành xáo các dữ liệu và tách thành tập huấn luyện và tập kiểm thử hay ta nên lựa chọn tham số shuffle = True (Tức là để mặc định)