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 $D = \{(X, y)|X \in R^{m \times n} \text{ và } y \in R^m\}$ được chia thành 2, gồm:

- Tập dữ liệu huấn luyện D_{train} thông thường bằng 70% D;
- Tập dữ liệu kiểm thử D_{test} 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.

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.

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