

Nhập Môn Trí Tuệ Nhân Tạo

Khoa Học Dữ Liệu

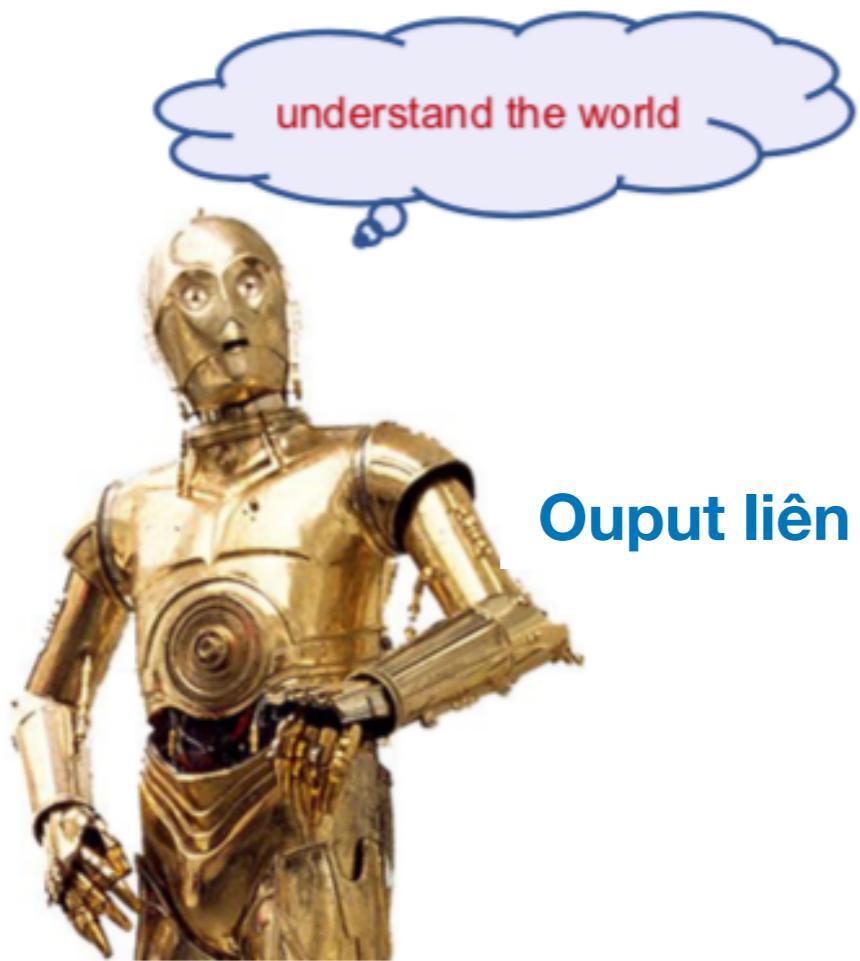
Anh-Tuan Nguyen
Tien-Lam Pham
Phenikaa School of Computing



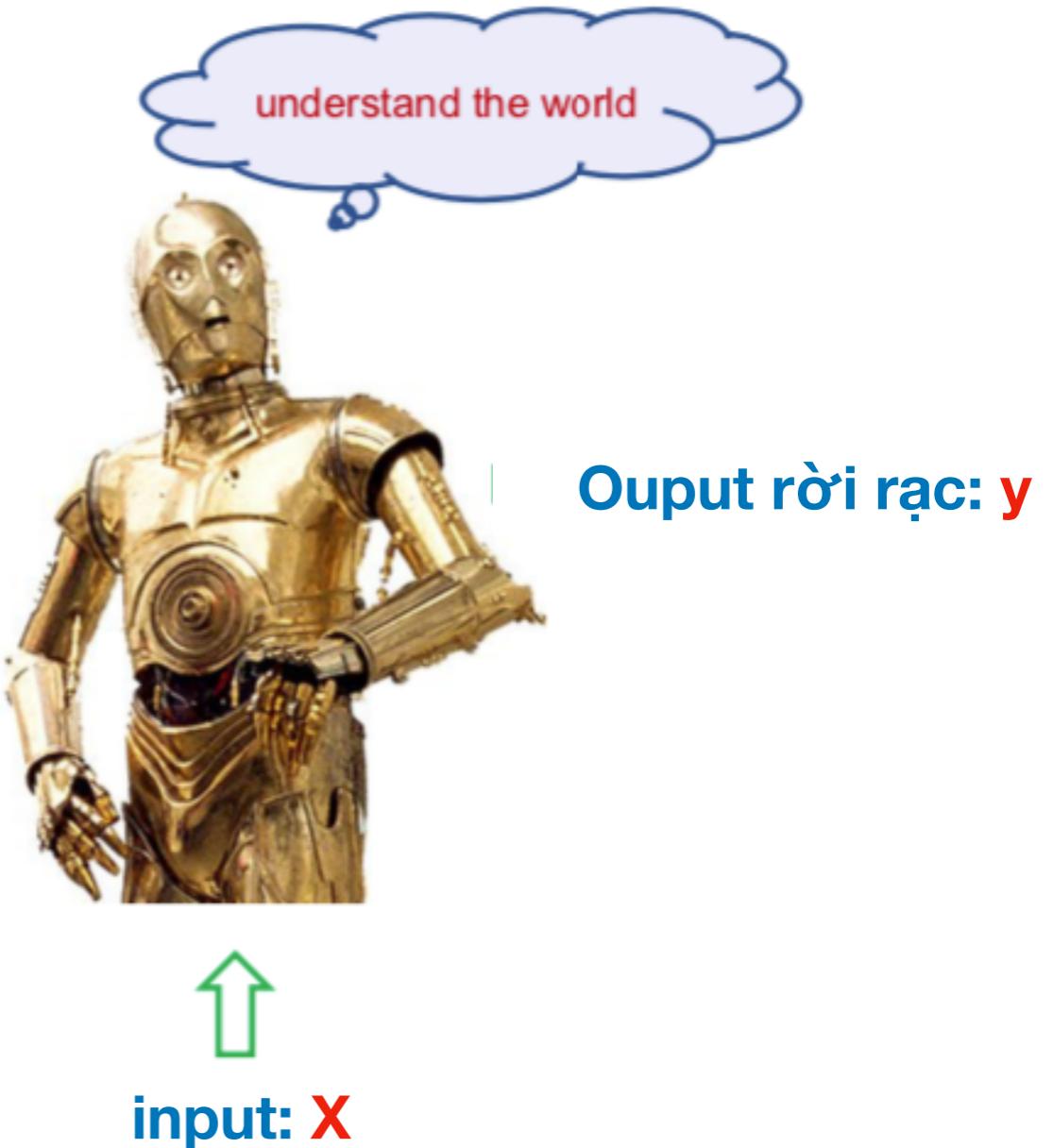
Data

- Population (quần thể)
- Tập mẫu
- Tập dữ liệu
- Feature vectors
- Đặc trưng thống kê: min, max, mean, variance, standardeviation
- Hệ số tương quan
- Phân tích biểu đồ: scatter plot, histogram

Bài toán phân loại (classification)



**Bài toán hồi qui
(regression)**



**Bài toán phân loại
(classification)**

Machine learning: Data-Driven Approach

(1) Problem setting

(2) Data collection

$$D = \{(x_i, y_i), i = 1, 2, \dots, m\}$$

(3) Modeling and Training Models

```
def train(images, labels):
    # Machine learning!
    return model
```

(4) Model selection

(5) Deploy suitable model (Using the best model to make prediction)

```
def predict(model, test_images):
    # Use model to predict labels
    return test_labels
```

Ví dụ

Dự đoán khách hàng lựa chọn 1 sản phẩm

- Input thông tin của khách hàng
- Output = 1 nếu khách hàng lựa chọn sản phẩm, Output = 0 nếu khách hàng không lựa chọn

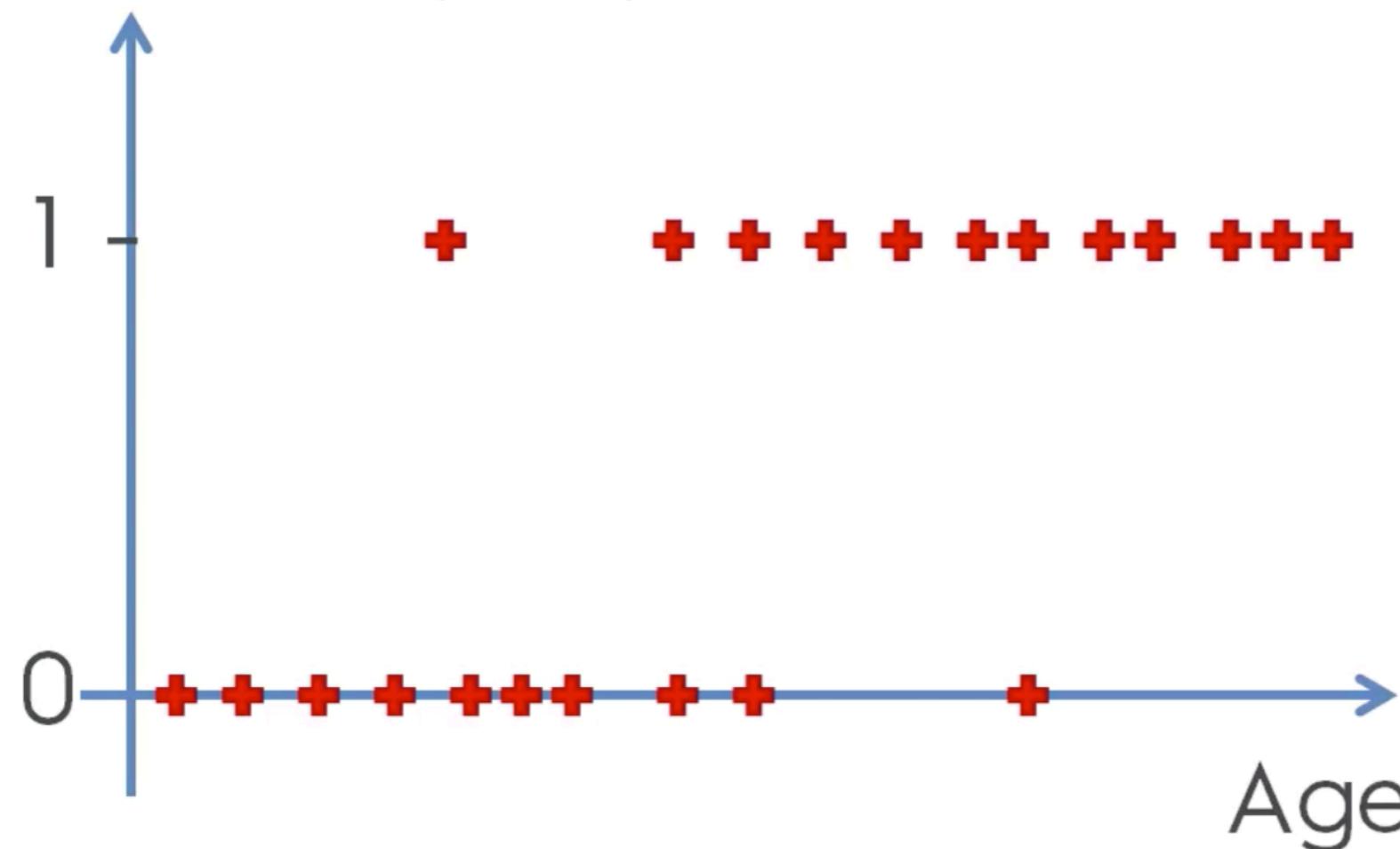
Data Driven Approach

- (1) Thu thập dữ liệu khách hành
- (2) Xây dựng và huấn luyện mô hình dự đoán khả năng lựa chọn sản phẩm của khách hàng
- (3) Sử dụng mô hình để dự đoán khả năng (xác suất) lựa chọn sản phẩm của khách hàng của khách hàng

Ví dụ

Dự đoán khách hàng lựa chọn 1 sản phẩm

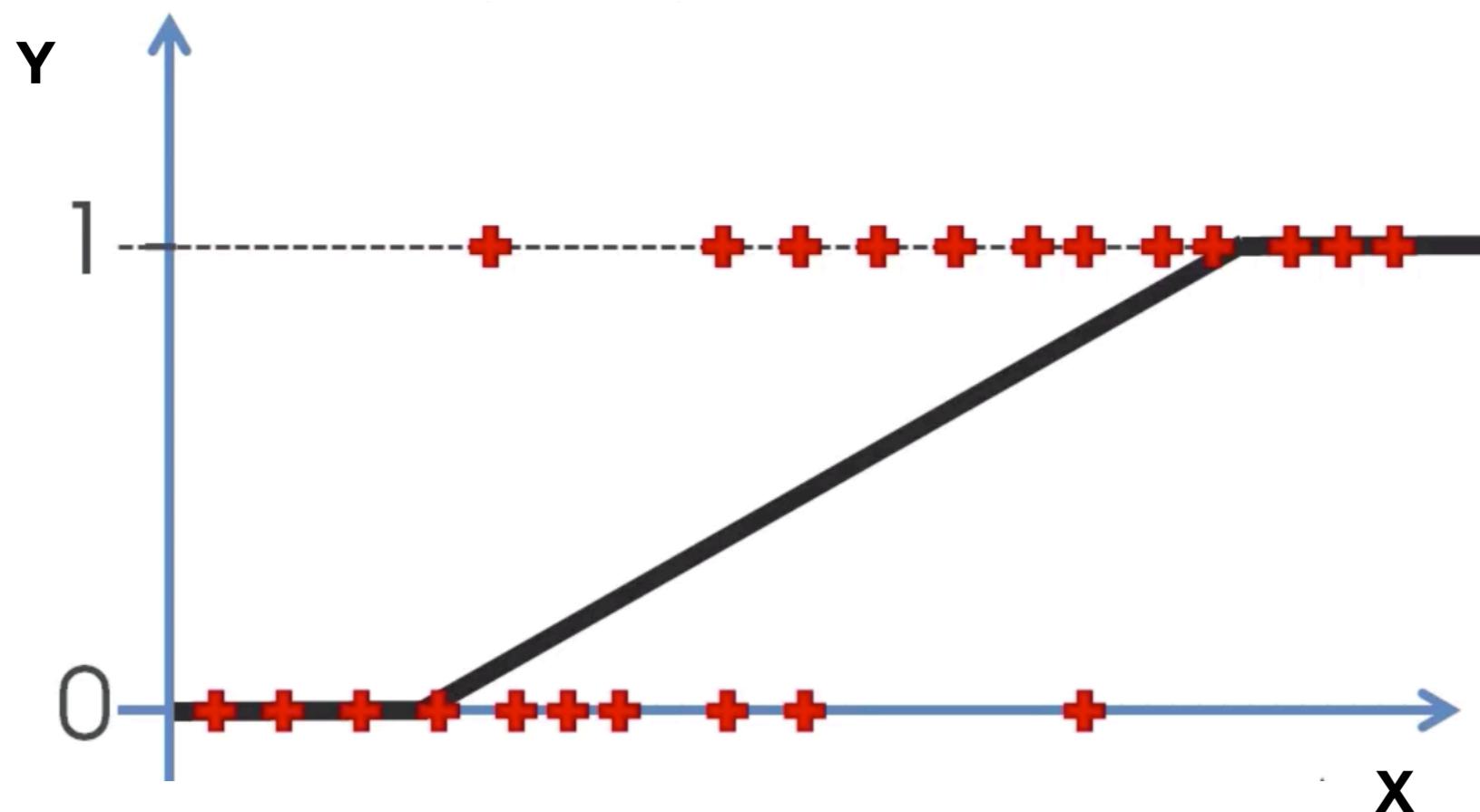
- Input thông tin của khách hàng
- **Output = 1** nếu khách hàng lựa chọn sản phẩm, **Output = 0** nếu khách hàng không lựa chọn



Ví dụ

Dự đoán khách hàng lựa chọn 1 sản phẩm

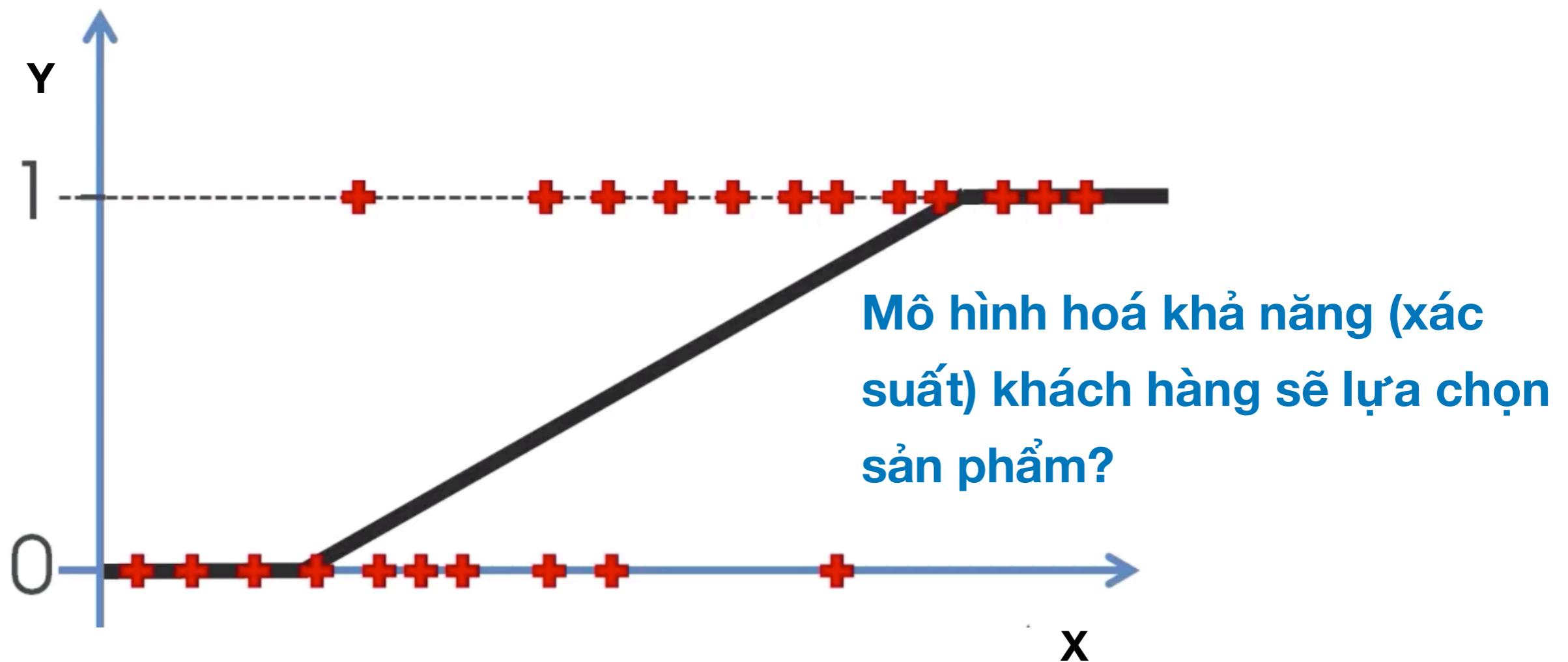
- Input thông tin của khách hàng
- **Output = 1** nếu khách hàng lựa chọn sản phẩm, **Output = 0** nếu khách hàng không lựa chọn



Logistic regression

Dự đoán khách hàng lựa chọn 1 sản phẩm

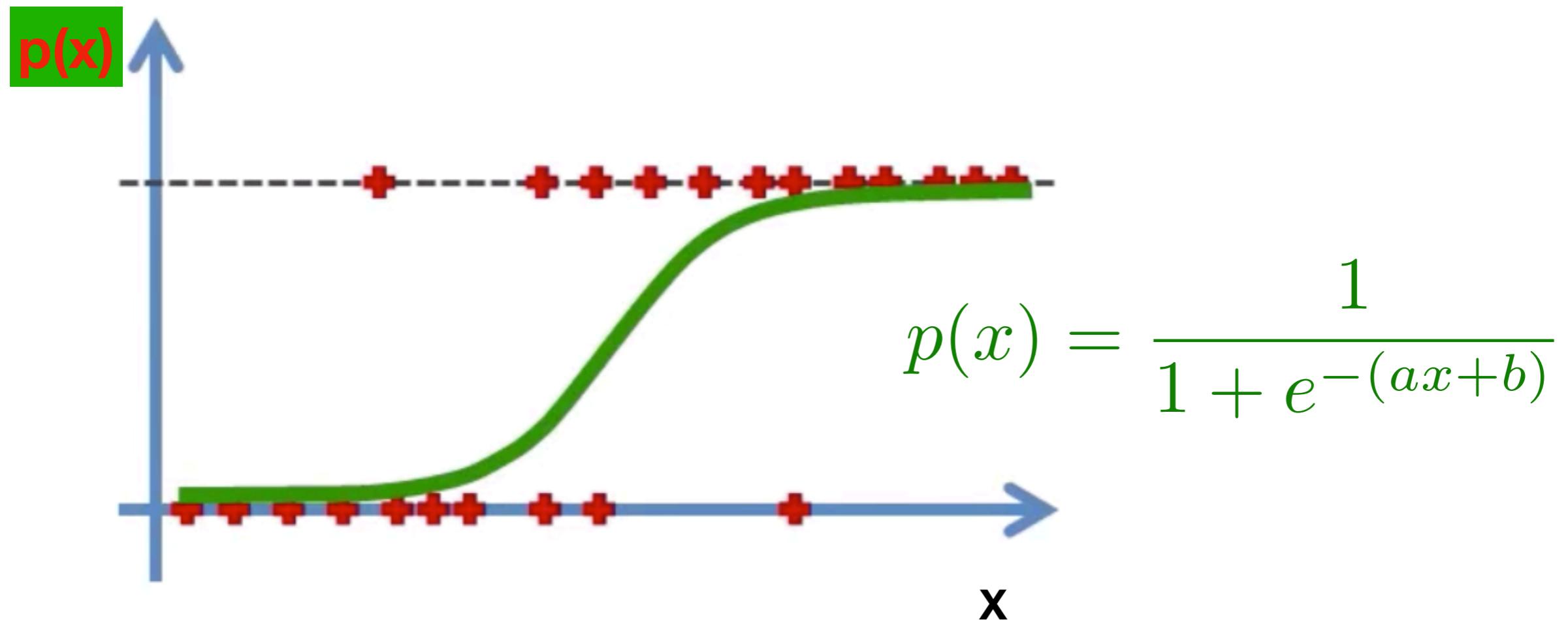
- Input thông tin của khách hàng
- **Output = 1** nếu khách hàng lựa chọn sản phẩm, **Output = 0** nếu khách hàng không lựa chọn



Logistic regression

Dự đoán khách hàng lựa chọn 1 sản phẩm

- Input thông tin của khách hàng
- Output = 1 nếu khách hàng lựa chọn sản phẩm, Output = 0 nếu khách hàng không lựa chọn

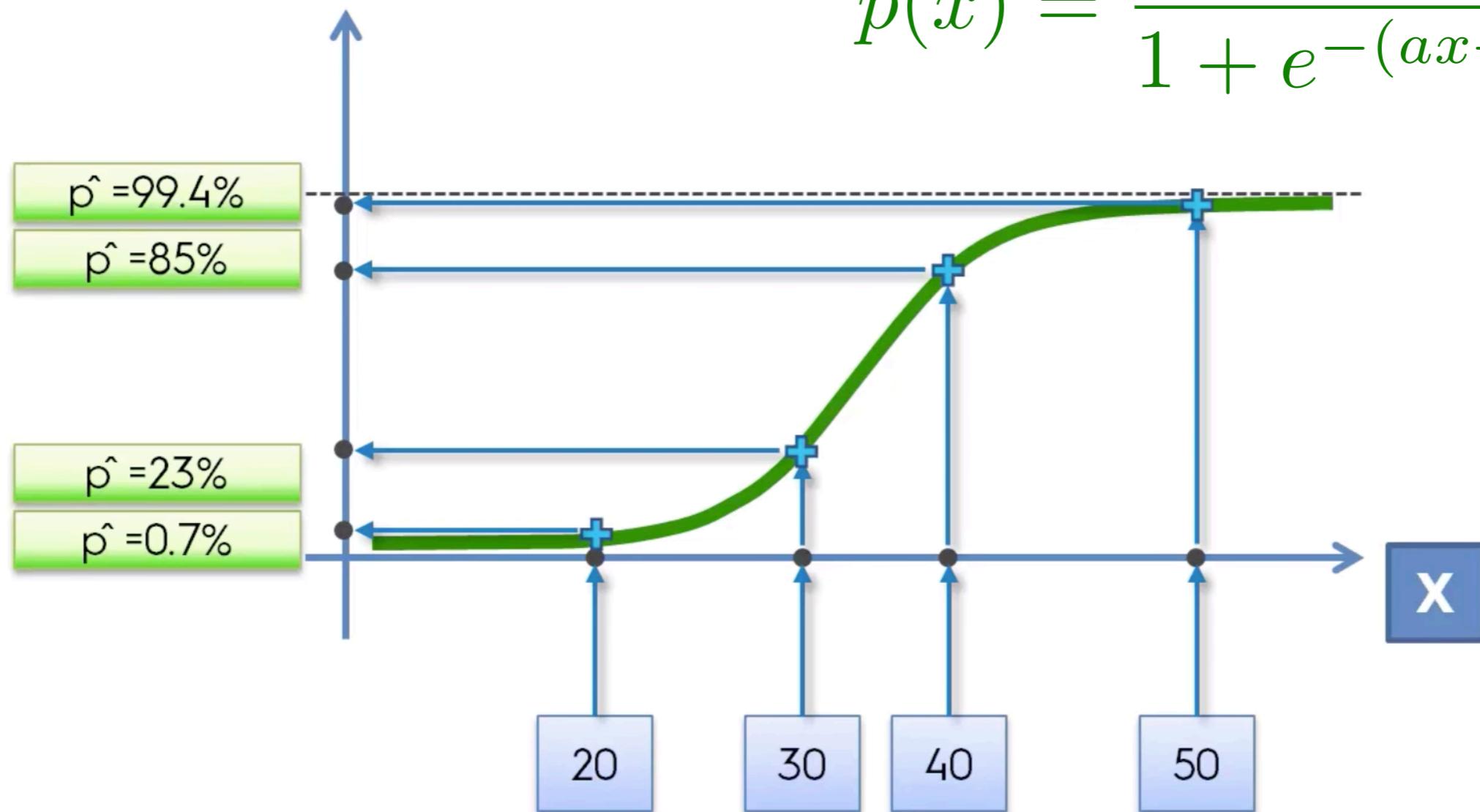


Logistic regression

Dự đoán khách hàng lựa chọn 1 sản phẩm

- Input thông tin của khách hàng
- **Output = 1** nếu khách hàng lựa chọn sản phẩm, **Output = 0** nếu khách hàng không lựa chọn

$$p(x) = \frac{1}{1 + e^{-(ax+b)}}$$

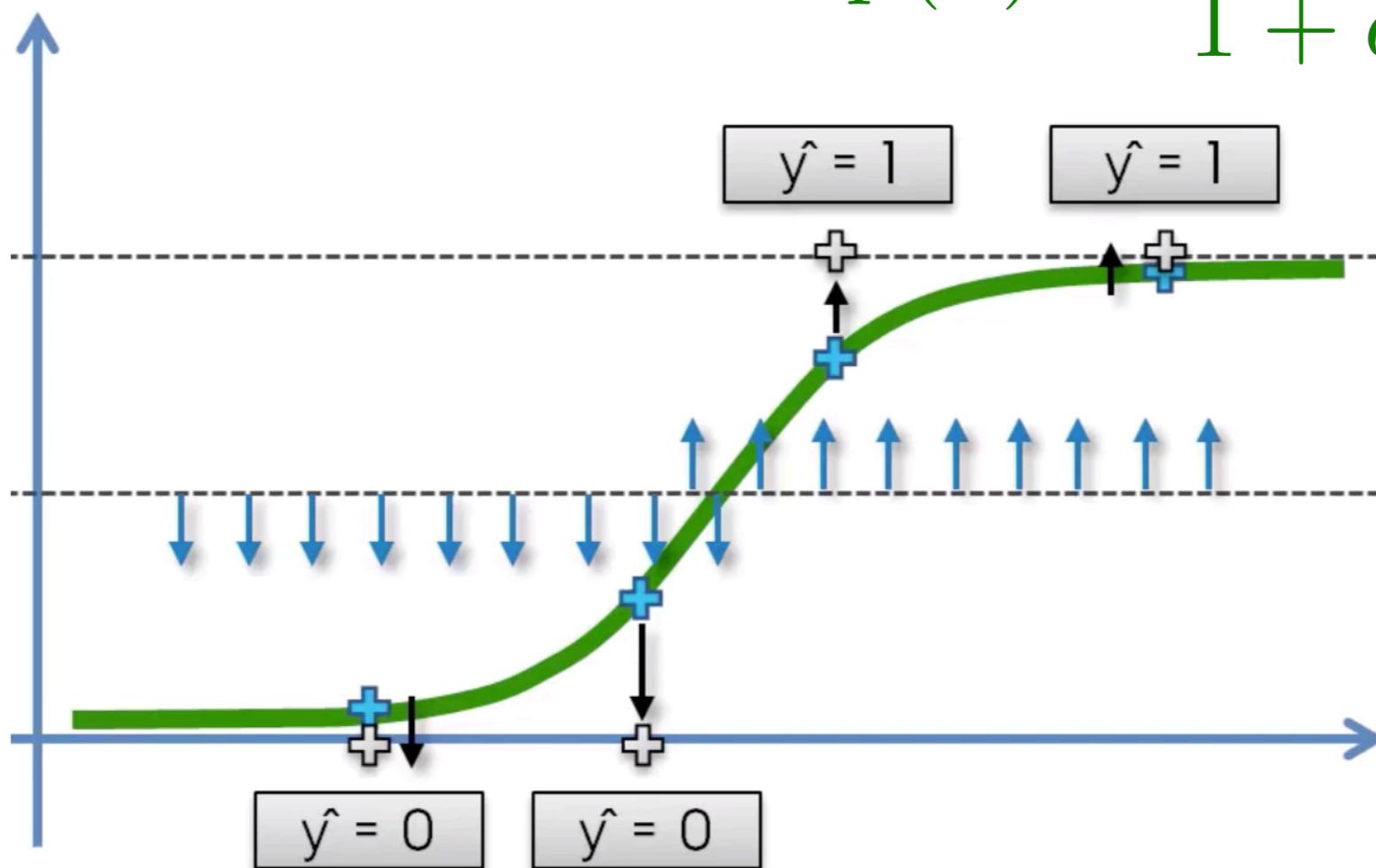


Logistic regression

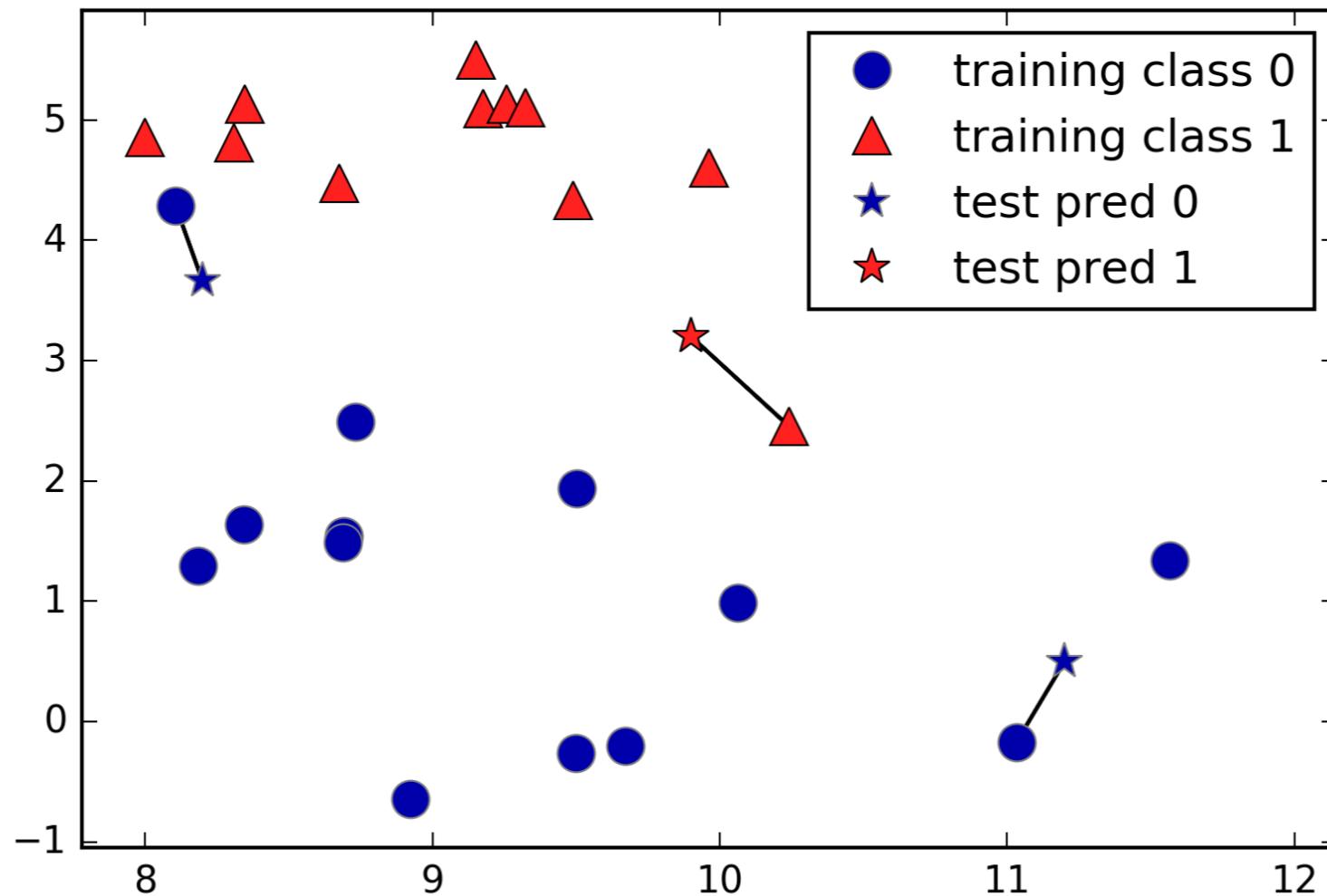
Dự đoán khách hàng lựa chọn 1 sản phẩm

- Input thông tin của khách hàng
- Output = 1 nếu khách hàng lựa chọn sản phẩm, Output = 0 nếu khách hàng không lựa chọn

$$p(x) = \frac{1}{1 + e^{-(ax+b)}}$$

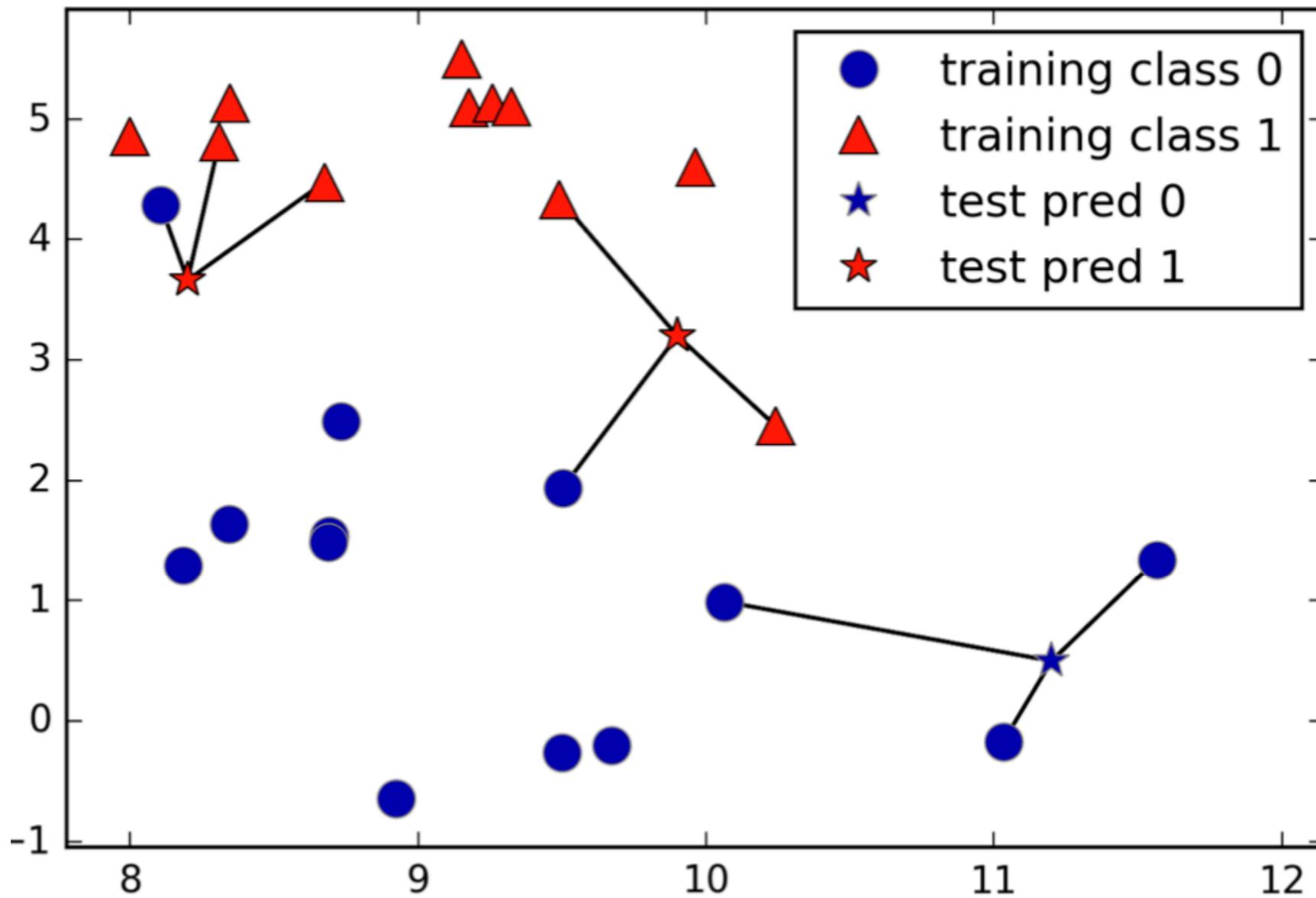


K-nearest classification



Dự đoán nhãn của các điểm dữ liệu test bằng mô hình **1-nearest neighbor**

K-nearest classification

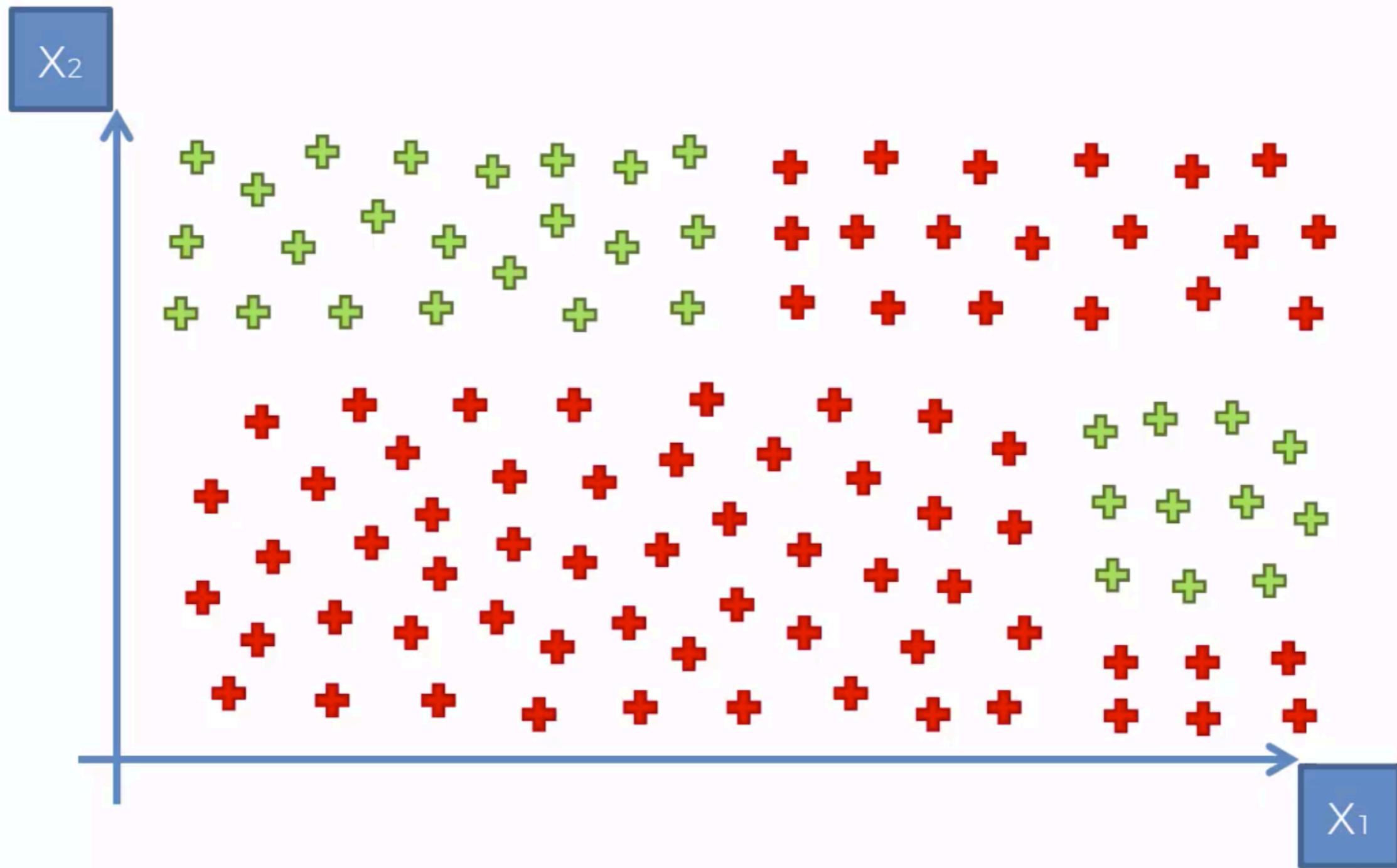


Dự đoán nhãn của các điểm dữ liệu test bằng mô hình **3-nearest neighbor**

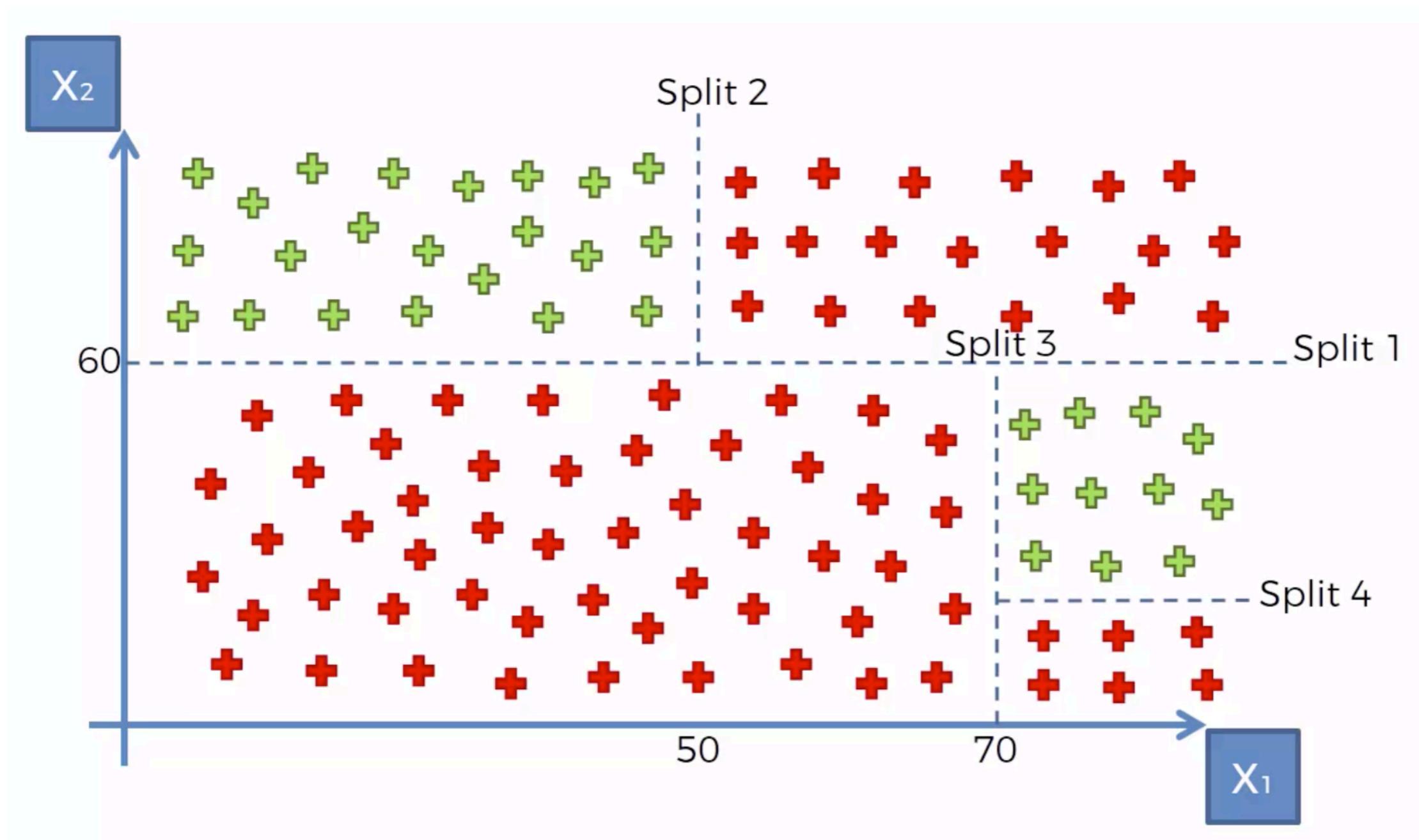
K-nearest classification

- (1) Chọn số neighbors (K)**
- (2) Xác định K điểm gần nhất của điểm dữ liệu mới cần dự đoán**
- (3) Trong K điểm dữ liệu gần nhất, đếm số điểm dữ liệu thuộc mỗi lớp**
- (4) Điểm dữ liệu mới sẽ thuộc loại chiếm ưu thế trong K điểm dữ liệu gần nhất**

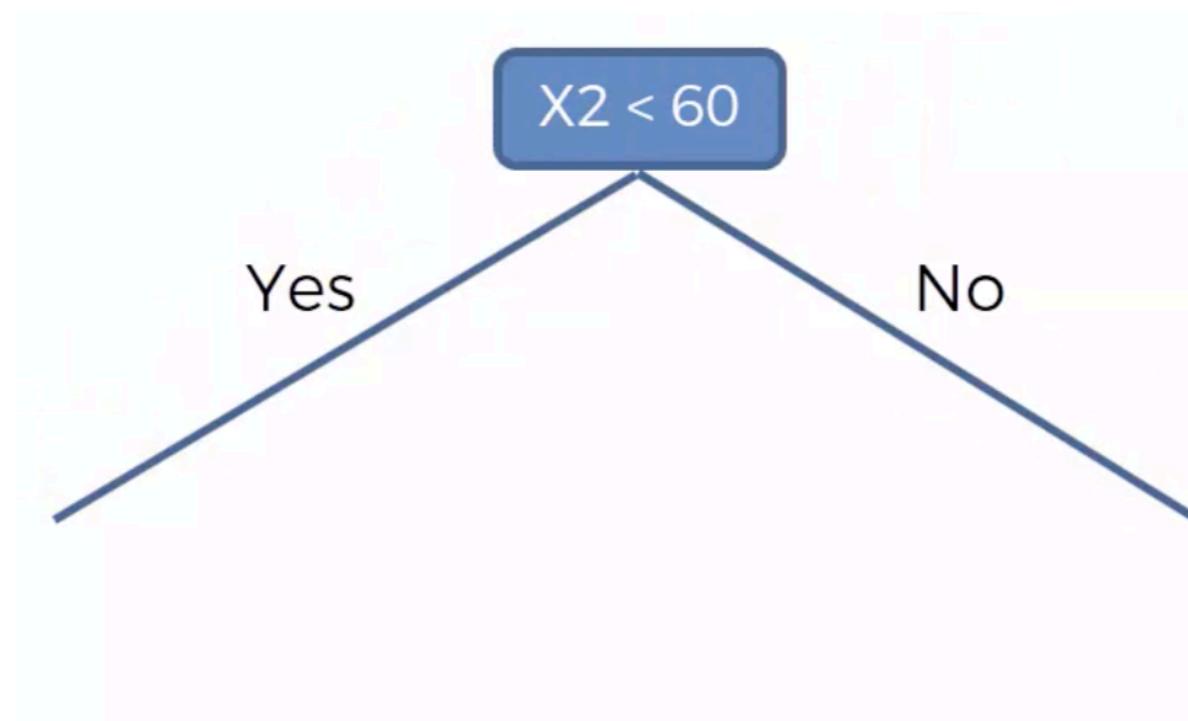
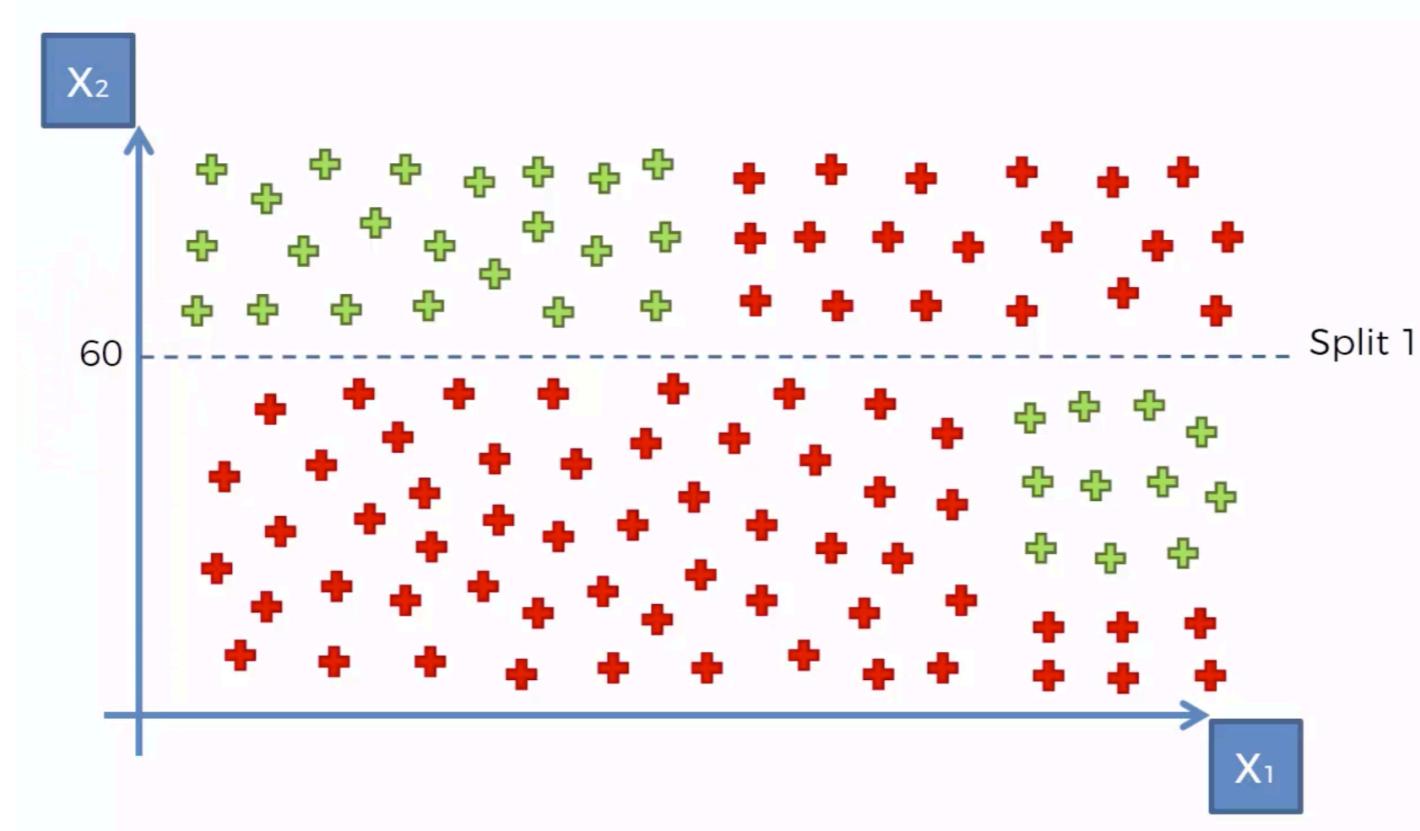
Cây quyết định (Decision tree)



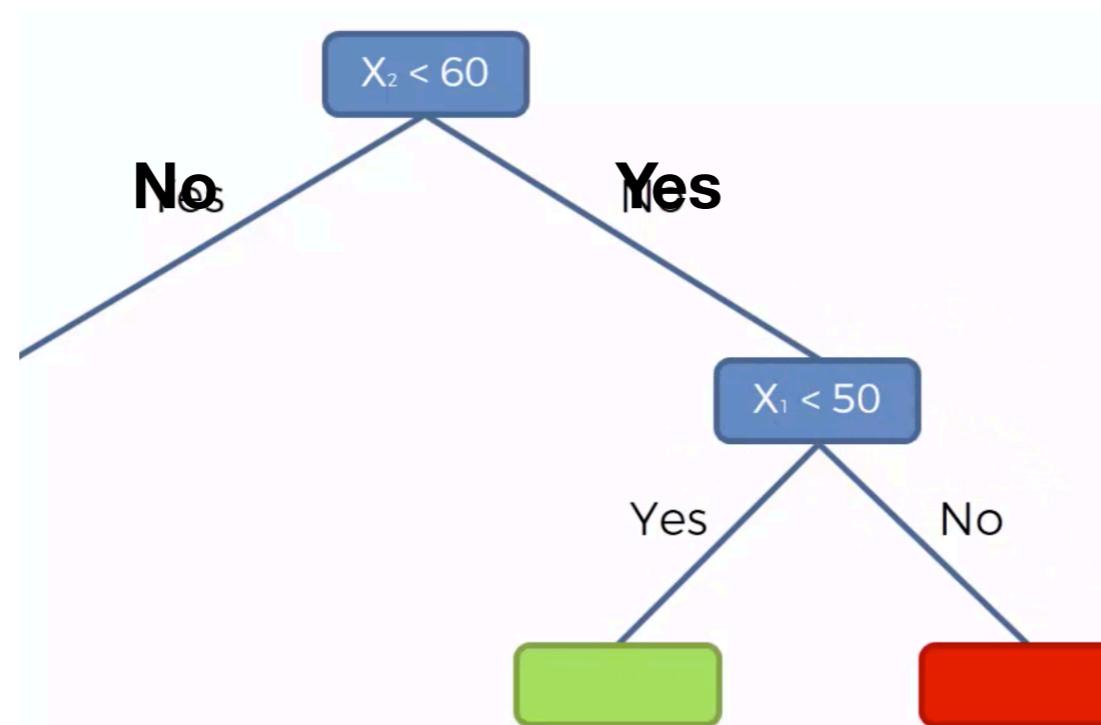
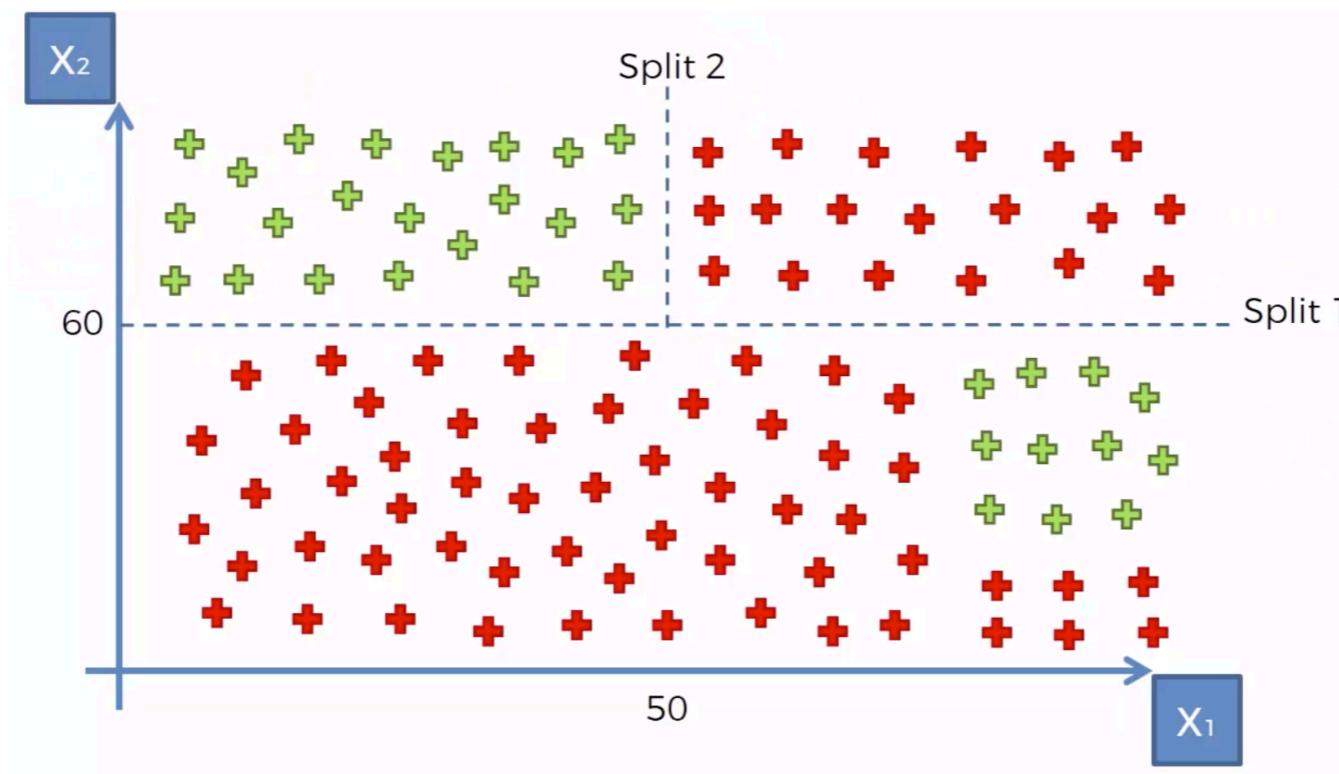
Cây quyết định (Decision tree)



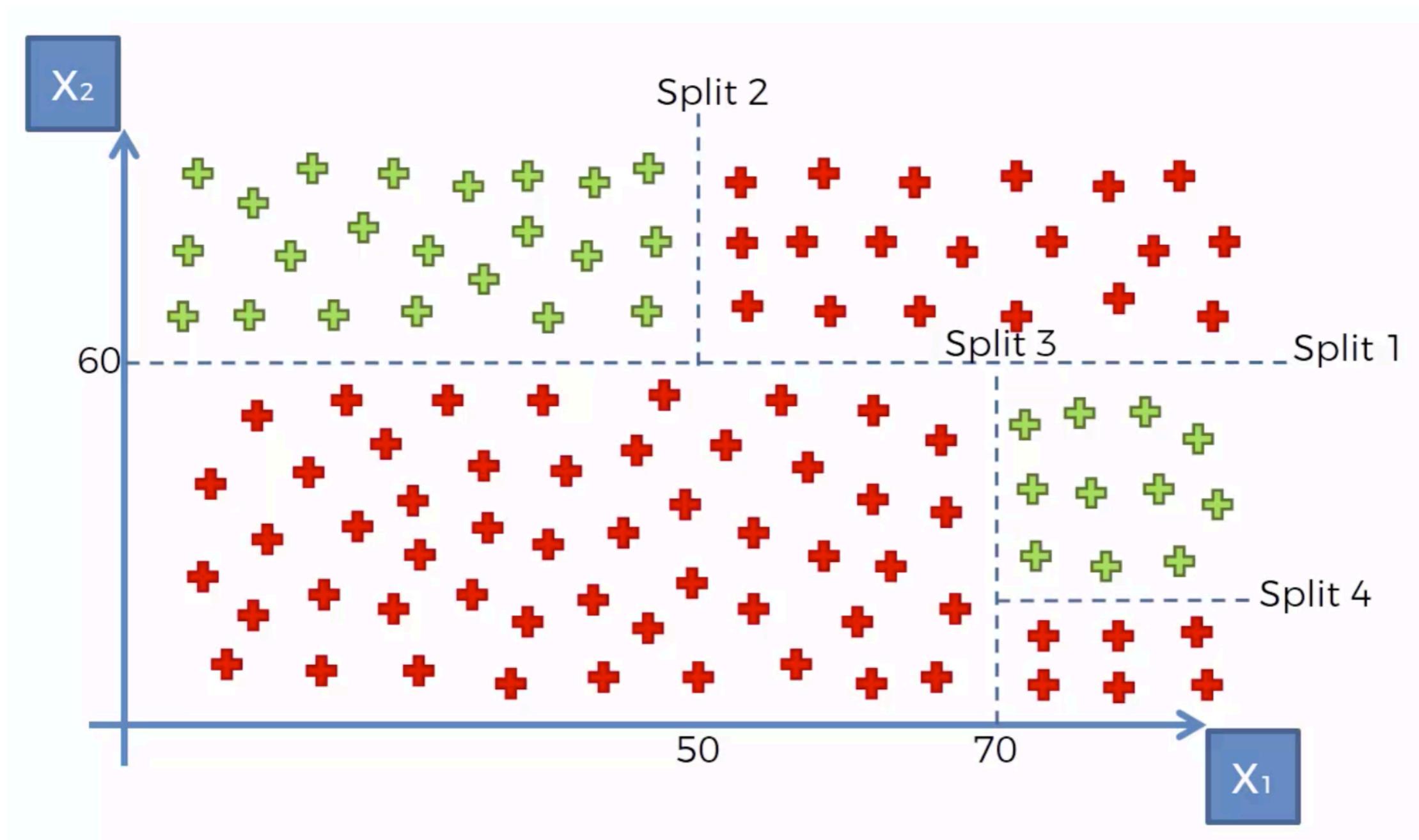
Cây quyết định (Decision tree)



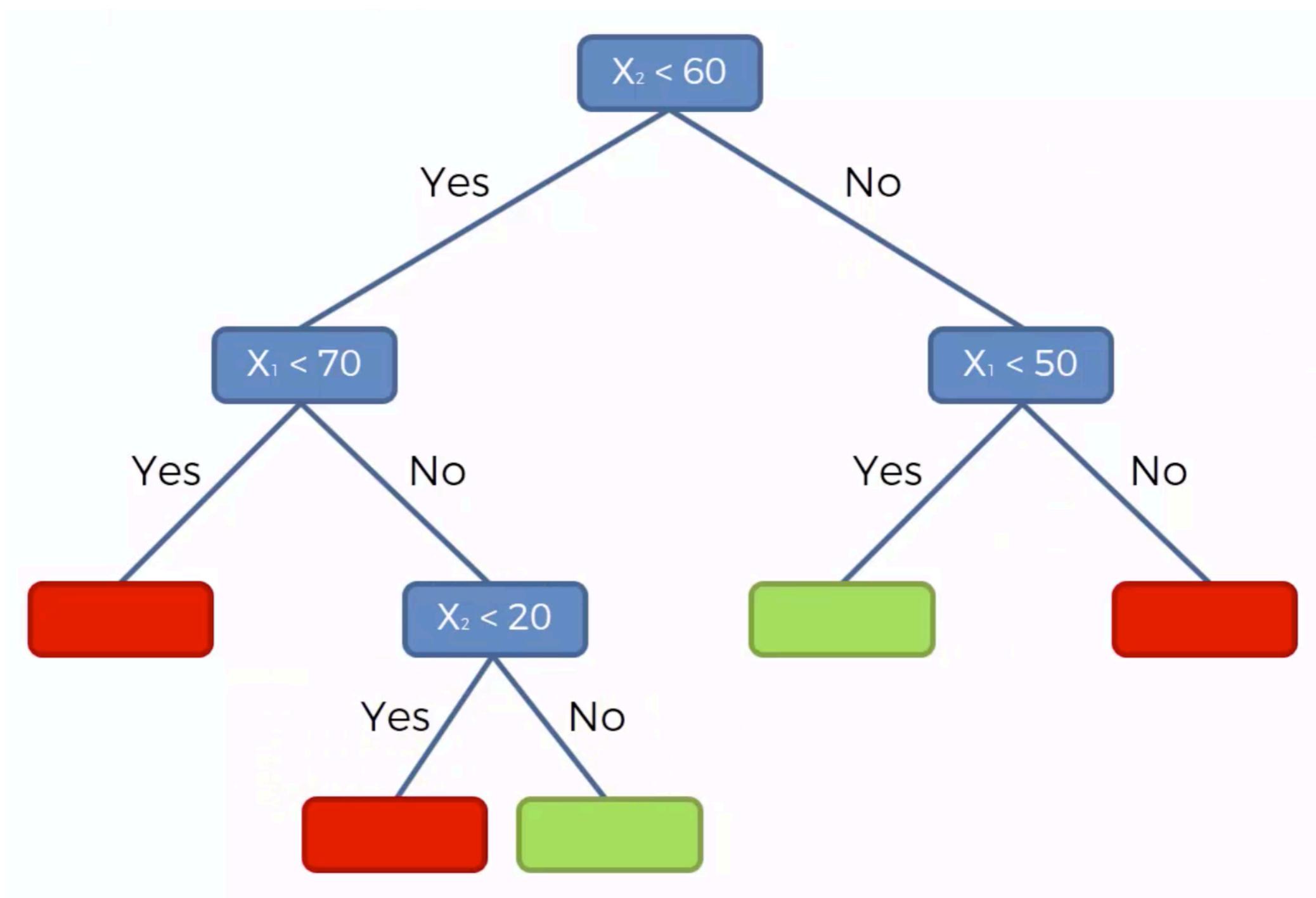
Cây quyết định (Decision tree)



Cây quyết định (Decision tree)



Cây quyết định (Decision tree)



Bài tập

- Một hãng ô tô khảo sát khả năng mua một loại xe mới của khách hàng bằng việc điều tra các thông tin của khách hàng thông qua mạng xã hội.
- Thông qua mạng xã hội hãng sẽ thu thập được thông tin về tuổi (age) của khách hàng, và có thể đánh giá được thu nhập của khách hàng.
- Hãng muốn biết những khách hàng như thế nào thì sẽ có khả năng mua xe ?

Bài tập

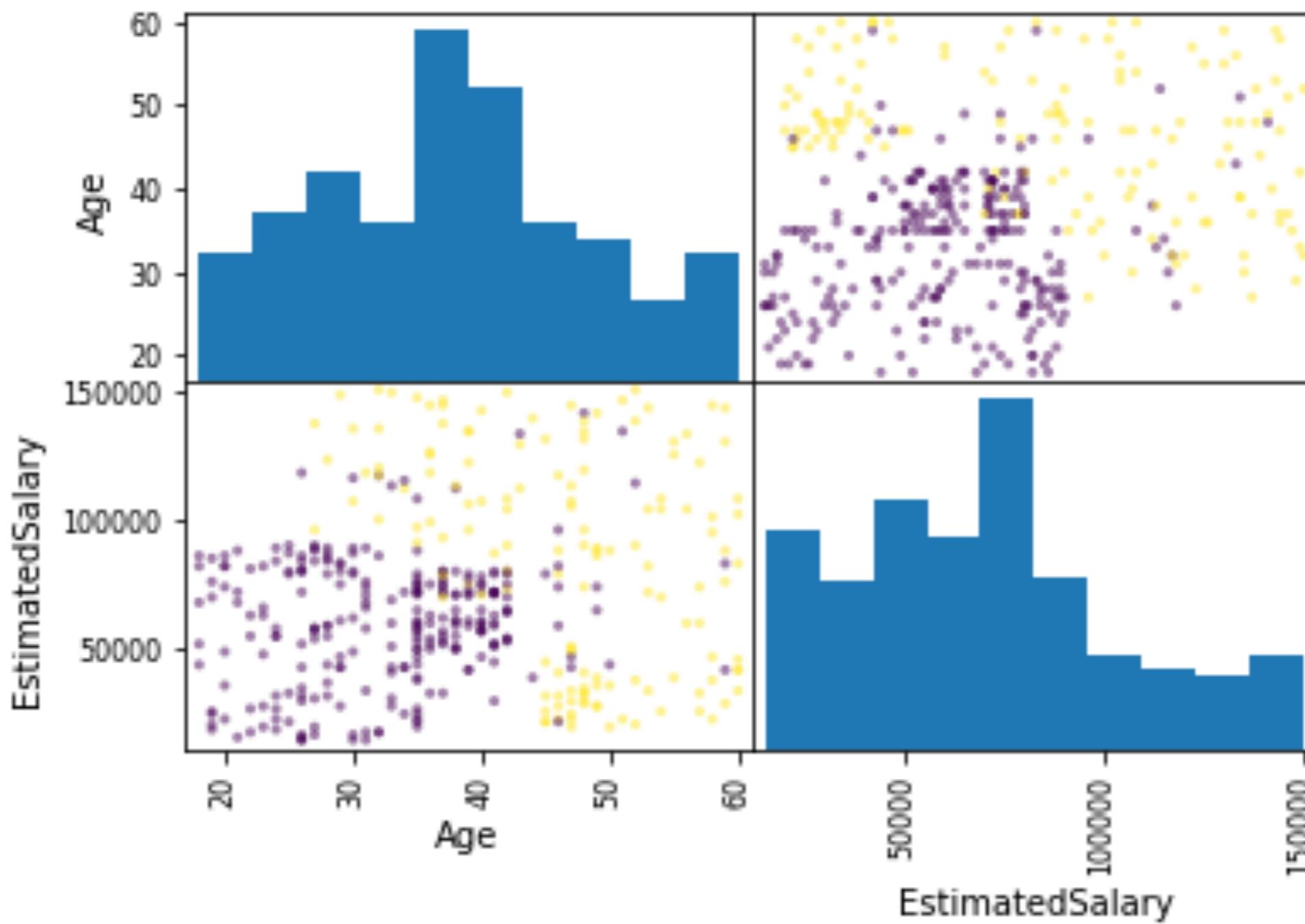
Data Driven Approach

- (1) Thu thập dữ liệu khách hành**
- (2) Xây dựng và huấn luyện mô hình dự đoán khả năng lựa chọn sản phẩm của khách hàng**
- (3) Sử dụng mô hình để dự đoán khả năng (xác suất) lựa chọn sản phẩm của khách hàng của khách hàng**

Bài tập

Index	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0
10	15570769	Female	26	80000	0
11	15606274	Female	26	52000	0
12	15746139	Male	20	86000	0

Bài tập



Bài tập



Bài tập thực hành

- From pandas import ‘read_csv’
- Đọc dataframe từ .csv: df = read_csv(“file_name”)
- Tính trung bình thu nhập của các khách hàng
- Tính độ tuổi trung bình của khách hàng
- Vẽ histogram thu nhập của khách hàng
- Tách trường giới tính ra và lưu vào numpy array
- Tính tỉ lệ nam, nữ
- Dùng hàm plt.bar vẽ bar plot với chiều cao là tỉ lệ nam nữ
- Dùng hàm OnehotEncoder của sklearn để chuyển array giới tính thành dạng số
- Dùng numpy array X để lưu thông tin khách hàng: cột đầu là tuổi, cột thứ 2 là thu nhập
- Dùng numpy array y để lưu thông tin mua hay không mua của khách hàng
- Vẽ scatter plot của X, điểm màu đỏ là khách hàng mua, màu xanh là khách hàng không mua
- Dùng hàm train_test_split để tạo X_train, X_test, y_train, y_test từ X và y
- Dùng hàm LogisticRegression để xây dựng mô hình dự đoán xác suất mua hàng của khách hàng