

**Giảng viên hướng dẫn:** Dương Việt Hằng

**Sinh viên thực hiện:**

Nguyễn Trọng Nhân 22521005

**TP. Hồ Chí Minh, tháng 10 năm 2023**

**BÁO CÁO BÀI TẬP 1**

**TOÁN CHO KHOA HỌC MÁY TÍNH**

**CS115.O12**

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Mục lục

[I. Giới thiệu bài toán Perceptron 2](#_Toc147228800)

[1. Giới thiệu 2](#_Toc147228801)

[2. Phát biểu bài toán 2](#_Toc147228802)

[II. Thuật toán Perceptron (Perceptron Learning Algorithm) 3](#_Toc147228803)

[1. Xây dựng hàm mất mát 3](#_Toc147228804)

[2. Tóm tắt thuật toán Perceptron 6](#_Toc147228805)

[3. Chứng minh hội tụ 6](#_Toc147228806)

[III. Minh hoạ với Python 6](#_Toc147228807)

[1. Tải bộ dữ liệu 6](#_Toc147228808)

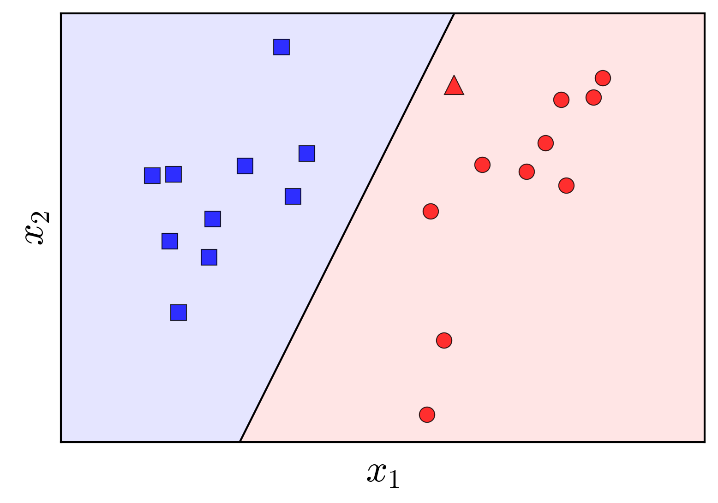
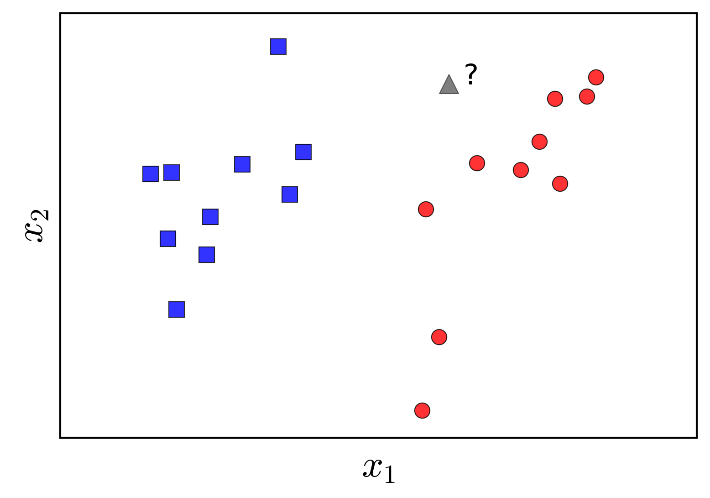
[2. Minh hoạ PLA 8](#_Toc147228809)

# Giới thiệu bài toán Perceptron

## Giới thiệu

Perceptron là một thuật toán phân lớp (classification) cho trường hợp đơn giản nhất: chỉ có hai lớp (bài toán với chỉ hai lớp được gọi là binary classification) và cũng chỉ hoạt động được trong một trường hợp rất cụ thể. Tuy nhiên, nó là nền tảng cho một mảng lớn quan trọng của Machine Learning là Neural Networks và sau này là Deep Learning.

Giả sử chúng ta có hai tập hợp dữ liệu đã được gán nhãn được minh hoạ trong hình 1 bên trái dưới đây. Hai lớp của chúng ta là tập các điểm màu xanh và tập các điểm màu đỏ. Bài toán đặt ra là: từ dữ liệu của hai tập được gán nhãn cho trước, hãy xây dựng một bộ phân lớp để khi có một điểm dữ liệu hình tam giác màu xám mới, ta có thể dự đoán được màu (nhãn) của nó.



Hình I‑1. Bài toán Perceptron

Hiểu theo một cách khác, chúng ta cần tìm lãnh thổ của mỗi lớp sao cho, với mỗi một điểm mới, ta chỉ cần xác định xem nó nằm vào lãnh thổ của lớp nào rồi quyết định nó thuộc lớp đó. Để tìm lãnh thổ của mỗi lớp, chúng ta cần đi tìm biên (boundary) giữa hai lãnh thổ này. Vậy bài toán phân lớp có thể coi là bài toán đi tìm biên giữa các lớp. Và biên đơn giản nhất trong không gian hai chiều là một đường thẳng, trong không gian ba chiều là một mặt phẳng, trong không gian nhiều chiều là một siêu mặt phẳng (hyperplane). Những biên phẳng này được coi là đơn giản vì nó có thể biểu diễn dưới dạng toán học bằng một hàm số đơn giản có dạng tuyến tính (linear). Tất nhiên, chúng ta đang giả sử rằng tồn tại một đường phẳng để có thể phân định lãnh thổ của hai lớp. Hình 1 bên phải minh họa một đường thẳng phân chia hai phân trong mặt phẳng. Phần có nền màu xanh được coi là lãnh thổ của lớp xanh, phần có nên màu đỏ được coi là lãnh thổ của lớp đỏ. Trong trường hợp này, điểm dữ liệu mới hình tam giác được phân vào lớp đỏ.

## Phát biểu bài toán

Bài toán Perceptron được phát biểu như sau: *Cho hai class được gán nhãn, hãy tìm một đường phẳng sao cho toàn bộ các điểm thuộc class 1 nằm về 1 phía, toàn bộ các điểm thuộc class 2 nằm về phía còn lại của đường phẳng đó. Với giả định rằng tồn tại một đường phẳng như thế.*

Nếu tồn tại một đường phẳng phân chia hai lớp thì ta gọi hai lớp đó là linearly separable. Các thuật toán phân lớp tạo ra các biên là các siêu mặt phẳng được gọi chung là Linear Classifier.

# Thuật toán Perceptron (Perceptron Learning Algorithm)

Ý tưởng cơ bản của PLA là xuất phát từ một nghiệm dự đoán nào đó, qua mỗi vòng lặp, nghiệm sẽ được cập nhật tới một ví trí tốt hơn. Việc cập nhật này dựa trên việc giảm giá trị của một hàm mất mát nào đó.

## Xây dựng hàm mất mát

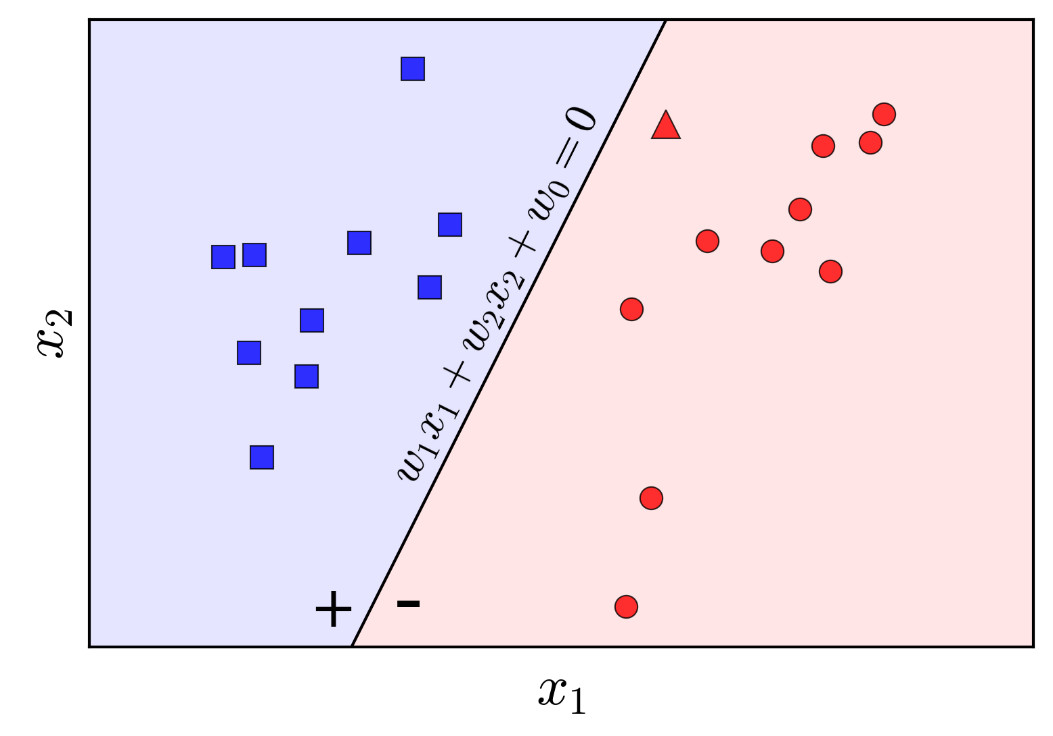
Giả sử là ma trận chứa các điểm dữ liệu mà mỗi cột là một điểm dữ liệu trong không gian chiều.

Các nhãn tương ứng với từng điểm dữ liệu được lưu trong một vector hàng  
, với (xanh) nếu thuộc lớp 1 và nếu thuộc lớp 2 (đỏ).

Tại một thời điểm, giả sử ta tìm được biên là đường phẳng có phương trình:

với là điểm dữ liệu mở rộng bằng cách thêm phần tử lên trước vector tương tự như trong hồi quy tuyến tính. Gọi là điểm dữ liệu mở rộng.

Với trường hợp , giả sử đường thẳng chính là nghiệm cần tìm như hình dưới đây:

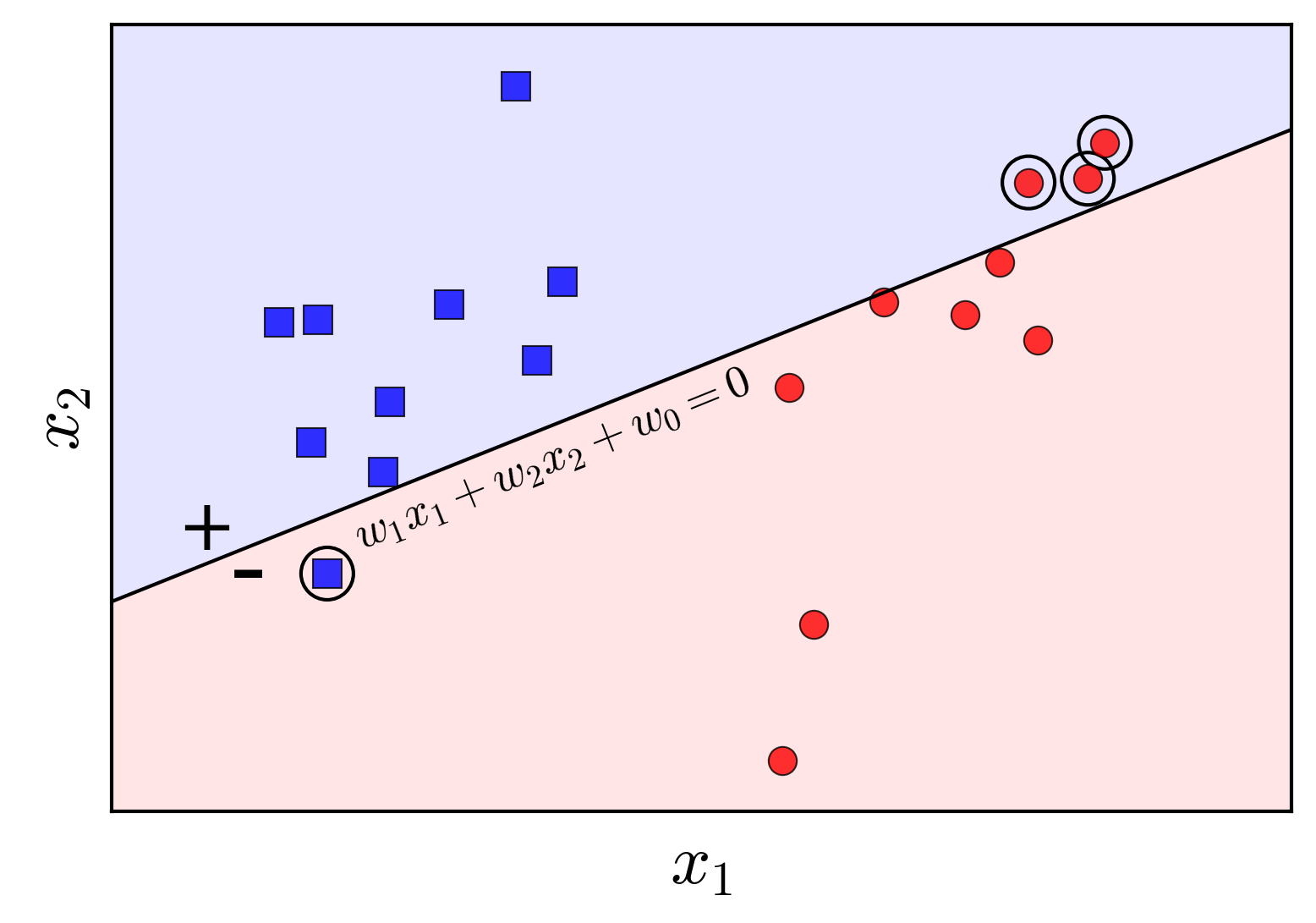


Hình II‑1. Phương trình đường thẳng biên

Nhận xét: Các điểm nằm về cùng một phía so với đường thẳng này sẽ làm cho hàm số mang cùng dấu. Giả sử các điểm nằm trong nửa mặt phảng nề xanh mang dấu , còn lại mang dấu . Các dấu này cũng tương đương với nhãn của mỗi lớp. Vậy nếu là một nghiệm của bài toán Perceptron, với một điểm dữ liệu mới chưa được gán nhãn, ta có thể xác định lớp của nó bằng phép toán:

Trong đó, là hàm xác định dấu, với giả sử .

Tiếp theo ta cần xây dựng hàm mất mát với tham số bất kỳ. Vẫn trong không gian hai chiều, giả sử đường thẳng được cho như hình dưới đây:



Hình II‑2. Đường thẳng bất kỳ và các điểm phân lớp sai được khoanh tròn.

Các điểm được khoanh tròn là các điểm bị phân lớp sai. Hàm mất mát đơn giản nhất mà ta nghĩ đến là hàm đếm số lượng các điểm vị phân lớp sai và tìm các cực tiểu hoá hàm này:

trong đó ℳ là tập hợp các điểm bị phân lớp sai. Với mỗi , vì điểm này bị đánh dấu sai (ngược với ) nên . Vậy chính là hàm đếm số lượng các điểm bị phân lớp sai. Khi hàm này bằng 0 thì ta không còn điểm nào bị phân lớp sai.

Hàm này là một hàm rời rạc, không tính được đạo hàm theo nên rất khó tối ưu. Chúng ta cần tìm một hàm mất mát khác để việc tối ưu hoá khả thi hơn.

Xét hàm mất mát sau:

Với việc bỏ đi hàm , một điểm bị phân lớp sai càng nằm xa biên thì giá trị càng lớn. Giá trị nhỏ nhất của hàm này là 0 nếu không có điểm nào bị phân lớp sai. Hàm mất mát này cũng được cho là tốt hơn vì nó trừng phạt các điểm lấn sâu trong lãnh thổ của lớp kia. Trong khi trừng phạt các điểm bị phân lớp sai như nhau, bất kể xa hay gần biên.

Tại một thời điểm, nếu chúng ta chỉ quan tâm tới các điểm bị misclassified thì hàm số khả vi, vậy chúng ta có thể sử dụng Gradient Descent hoặc Stochastic Gradient Descent (SGD) để tối ưu hàm mất mát này. Với ưu điểm của SGD cho các bài toán large-scale, chúng ta sẽ làm theo thuật toán này. Với mỗi điểm dữ liệu bị phân lớp sai, hàm mất mát trở thành:

Đạo hàm tương ứng:

Vậy quy tác cập nhật là:

với η là tốc độ học (learning rate) được chọn bằng 1. Ta có quy tắc cập nhật:

Nói cách khác, với mỗi điểm bị phân lớp sai, ta chỉ cần nhân nó với nhãn của nó, cộng vào để được mới.

Ta có:

Nếu , vì bị phân lớp sai nên . Cũng vì nên  
, nghĩa là , hay tiến về phía làm cho được phân lớp đúng. Điều tương tự xảy ra nếu .

## Tóm tắt thuật toán Perceptron

1. Chọn ngẫu nhiên một vector hệ số với các phần tử gần 0
2. Duyệt ngẫu nhiên qua từng điểm dữ liệu

* Nếu được phân lớp đúng, chúng ta không cần làm gì
* Nếu bị phân lớp sai, cập nhật theo công thức:

1. Kiểm tra xem còn bao nhiêu điểm bị phân lớp sai. Nếu không còn, ta kết thúc thuật toán. Nếu còn, quay lại bước 2.

## Chứng minh hội tụ

Giả sử rằng là một nghiệm của bài toán (ta có thể giả sử việc này được vì chúng ta đã có giả thiết hai lớp là linearly separable - tức tồn tại nghiệm). Có thể thấy rằng, với mọi , nếu là nghiệm,  cũng là nghiệm của bài toán. Xét dãy số không âm . Với là một điểm bị phân lớp sai, nếu dùng nghiệm ta có:

Dấu nhỏ hơn ở dòng cuối là vì và . Nếu ta đặt:

và chọn , ta có:

Điều này nghĩa là: nếu luôn luôn có các điểm bị misclassified thì dãy là dãy giảm, bị chặn dưới bởi 0, và phần tử sau kém phần tử trước ít nhất một lượng là . Điều vô lý này chứng tỏ đến một lúc nào đó sẽ không còn điểm nào bị misclassified. Nói cách khác, thuật toán PLA hội tụ sau một số hữu hạn bước.

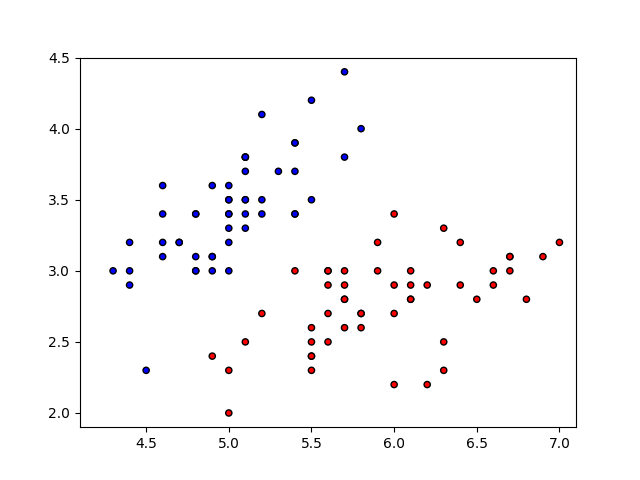
# Minh hoạ với Python

## Tải bộ dữ liệu

Dùng bộ dữ liệu iris dataset của thư viện scikit-learn phân loại hoa diên vĩ dựa trên chiều dài và chiều rộng của đài hoa (perceptron\_test.py):



Bộ dữ liệu gồm 3 lớp, mỗi lớp có 50 mục, dữ liệu mỗi mục gồm 4 số: chiều dài, chiều rộng của đài hoa và chiều dài, chiều rộng của cánh hoa. Ở bài toán này, ta chỉ cần 2 thông tin của đài hoa và 2 lớp được đánh nhãn 0 và 1 (loài setosa và versicolor) được đánh nhãn lại tương ứng là 1 và -1. Bộ dữ liệu được minh hoạ trong hình dưới đây:



Hình III‑1. Minh hoạ bộ dữ liệu. Hai nhãn 1 và -1 lần lượt được tô màu xanh và đỏ.

Các điểm xanh được đánh nhãn 1, các điểm đỏ được đánh nhãn -1. Do chỉ lấy 2 trong số 4 đặc trưng dữ liệu, lớp 1 có 40 điểm dữ liệu, lớp -1 có 44 điểm dữ liệu được minh hoạ do trùng lặp, không phải 50 điểm mỗi lớp như trong bộ dữ liệu.

## Minh hoạ PLA

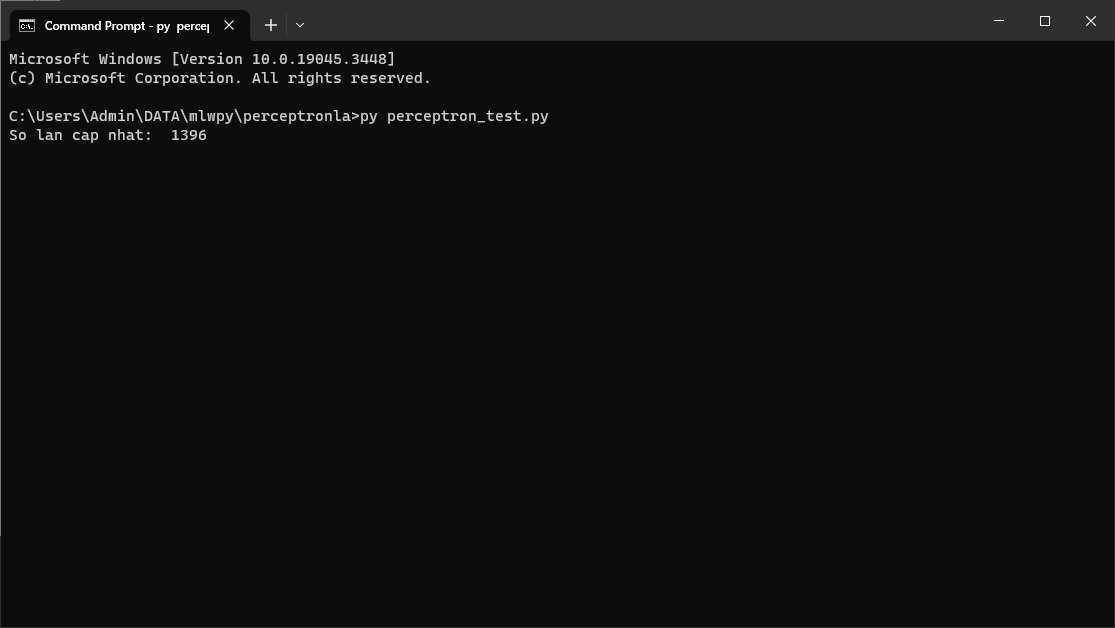
Dựa vào tóm tắt thuật toán đã nêu trên, ta có mã cho thuật toán PLA (perceptron.py):



Thực thi thuật toán và vẽ biểu đồ với bộ dữ liệu trong file perceptron\_test.py:

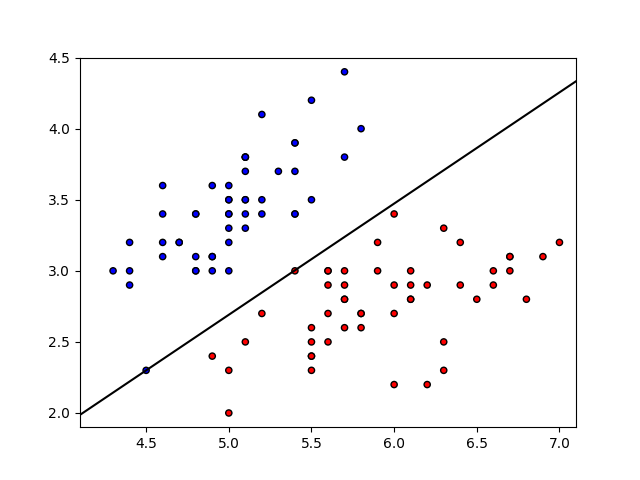


Kết quả thực thi:



Hình III‑2. Kết quả thực thi

Biểu đồ với dữ liệu được phân loại:



Hình III‑3. Dữ liệu đã được phân loại.