**Logistic Regression**

**Giới Thiệu:**

Mục tiêu của hồi qui Logistic là nghiên cứu mối tương quan giữa một (hay nhiều) yếu tố nguy cơ (risk factor) và đối tượng phân tích (outcome). Chẳn hạn như đối với nghiên cứu mối tương quan giữa thói quen hút thuốc lá và nguy cơ mắc ung thư phổi thì yếu tố nguy cơ ở đây là thói quen hút thuốc lá và đối tượng phân tích ở đây là nguy cơ mắc ung thư phổi. Trong hồi qui logistic thì các đối tượng nghiên cứu thường được thể hiện qua các biến số nhị phân (binary) như xảy ra/ không xảy ra ; chết/sống ; có/không,… còn các yếu tố nguy cơ có thể được thể hiện qua các biến số liên tục (tuổi, huyết áp,…) hoặc các biến nhị phân (giới tính) hay các biến thứ bậc (thu nhập : Cao, trung bình, thấp)

**Ví dụ:**

Một nhóm 20 sinh viên dành thời gian trong khoảng từ 0 đến 6 giờ cho việc ôn thi. Thời gian ôn thi này ảnh hưởng đến xác suất sinh viên vượt qua kỳ thi như thế nào?

Kết quả thu được như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Số giờ học | Kết quả | Số giờ học | Kết quả |
| 0.5 | 0 | 2.75 | 1 |
| 0.75 | 0 | 3 | 0 |
| 1 | 0 | 3.25 | 1 |
| 1.25 | 0 | 3.5 | 0 |
| 1.5 | 0 | 4 | 1 |
| 1.75 | 0 | 4.25 | 1 |
| 1.75 | 1 | 4.5 | 1 |
| 2 | 0 | 4.75 | 1 |
| 2.25 | 1 | 5 | 1 |
| 2.5 | 0 | 5.5 | 1 |

Mặc dù có một chút bất công khi học 3.5 giờ thì trượt, còn học 1.75 giờ thì lại đỗ, nhìn chung, học càng nhiều thì khả năng đỗ càng cao.

Biểu diễn các điểm này trên đồ thị:



*Hình 1: ví dụ về kết quả thi dựa trên số giờ học*

**Mô hình Logistic Regression**

Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng:

Trong đó: được gọi là logistic function. Một số activation cho mô hình tuyến tính được cho trong hình dưới đây:



*Hình 2: các activation khác nhau*

* **Đường màu vàng**:

Biểu diễn linear regression. Đường này không bị chặn nên không phù hợp cho bài toán này. Khi áp dụng mô hình linear regression như hình dưới đây và lấy mốc 0.5 để phân lớp, toàn bộ sinh viên thi trượt vẫn được dự đoán là trượt, nhưng rất nhiều sinh viên thi đỗ cũng được dự đoán là trượt (nếu ta coi điểm x màu xanh lục là ngưỡng cứng để đưa ra kết luận). Rõ ràng đây là một mô hình không tốt.



*Hình 3: mô tả Linear Regression không phù hợp*

* **Đường màu đỏ:**

Rõ ràng là khi sử dụng ngưỡng cứng ở đây vì dữ liệu ở đây không tách biệt tuyến tính (linearly separable).

* **Các đường màu xanh:**

Hai đường màu xanh lá và xanh dương hợp với bài toán của ta hơn, chúng có 1 vài tính chất quan trọng sau:

* Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng .
* Nếu coi điểm có tung độ là 1/2 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này khớp với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.
* Mượt (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.

**Sigmoid function**

được sử dụng nhiều nhất, vì nó bị chặn trong khoảng . Thêm nữa:

**;**

**Hyperbolic tangent (tanh) function**

Hàm số này nhận giá trị trong khoảng **,** với khoảng hàm có dạng: