Linear regression

Linear regression là thuật toán dùng để xấp xỉ đầu ra một hệ thống /đối tượng theo một tổ hợp tuyến tính của các đầu vào:

Y ≈ A.X + B

***Bài 1:***

Dân số thế giới qua các năm [https://en.wikipedia.org/wiki/World\_population\_estimates](https://www.google.com/url?q=https://en.wikipedia.org/wiki/World_population_estimates&sa=D&ust=1575097242356000)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| Population (billion) | 6.067 | 6.137 | 6.215 | 6.314 | 6.396 | 6.477 | 6.555 | 6.625 | 6.705 | 6.810 | 6.892 |

Nếu tốc độ tăng dân số thế giới không đổi theo từng năm thì quan hệ giữa dân số và năm sẽ có dạng một đường thẳng, hay:

Population = a. Year + b

***Bài 2:***

Xây dựng mô hình để ước tính lượng mỡ máu theo tuổi và cân nặng dựa vào bảng dữ liệu sau:

|  |  |  |
| --- | --- | --- |
| Weight (kilograms) | Age (Years) | Blood fat content (mg/l) |
| 84 | 46 | 354 |
| 73 | 20 | 190 |
| 65 | 52 | 405 |
| 70 | 30 | 263 |
| 76 | 57 | 451 |
| 69 | 25 | 302 |
| 63 | 28 | 288 |
| 72 | 36 | 385 |
| 79 | 57 | 402 |
| 75 | 44 | 365 |
| 27 | 24 | 209 |
| 89 | 31 | 290 |
| 65 | 52 | 346 |
| 57 | 23 | 254 |
| 59 | 60 | 395 |
| 69 | 48 | 434 |
| 60 | 34 | 220 |
| 79 | 51 | 374 |
| 75 | 50 | 308 |
| 82 | 34 | 220 |
| 59 | 46 | 311 |
| 67 | 23 | 181 |
| 85 | 37 | 274 |
| 55 | 40 | 303 |
| 63 | 30 | 244 |

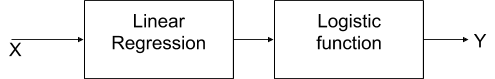
Logistic regression:

Logistic regression là thuật toán dùng để dự đoán một sự kiện có xảy ra hay không và xác suất xảy ra sự kiện là bao nhiêu.

Đầu vào của mô hình có thể có một hoặc nhiều biến số (tương tự Linear regression), đầu ra của mô hình là một giá trị nằm trong đoạn [0,1] thể hiện xác suất sẽ xảy ra sự kiện cần dự đoán.

***Cấu trúc của mô hình Logistic Regression:***

Mô hình Logistic Regression được xây dựng bằng cách ghép mô hình Linear Regression với hàm logistic:



Cấu trúc của mô hình logistic regression

Công thức toán học của hàm logistic:



Ý nghĩa của hàm logistic là tạo ra giá trị trong đoạn [0,1] ở đầu ra, do giá trị này thể hiện xác suất một sự kiện có thể xảy ra hay không.

Mối quan hệ giữa đầu ra và đầu vào của mô hình Logistic Regression:



***Bài 3:***

Ước tính khả năng qua bài thi cuối kì theo thời gian học ôn thi của sinh viên.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Số giờ học | 0.5 | 0.75 | 1 | 1.25 | 1.5 | 1.75 | 2 | 2.25 | 2.5 | 2.75 | 3 | 3.25 | 3.5 | 3.75 | 4 | 4.25 | 4.5 | 4.75 | 5 | 5.5 |
| Qua kì thi | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

***Bài 4:***

Dự đoán kết quả nộp hồ sơ cao học tại trường đại học UCLA.

Các sinh viên sau khi tốt nghiệp đại học có thể nộp hồ sơ để xin học tiếp chương trình cao học. Trong thông tin hồ sơ của một sinh viên, có các trường dữ liệu sau:

* GRE : điểm bài thi GRE chuyên ngành
* GPA : điểm trung bình đại học
* Rank : xếp hạng của trường đại học sinh viên đã học (cao nhất là 1, thấp nhất là 4)

Ví dụ về một số bộ dữ liệu trong hồ sơ của các sinh viên như sau:

|  |  |  |
| --- | --- | --- |
| gre | gpa | rank |
| 380 | 3.61 | 3 |
| 660 | 3.67 | 3 |
| 800 | 4.00 | 1 |
| 640 | 3.19 | 4 |
| 520 | 2.93 | 4 |
| 760 | 3.00 | 2 |