SUPERVISED LEARNING

Mô hình Hồi quy tuyến tính

TS. Trần Quang Quý

Tổng quan

Trong phần này trình bày các nội dung về các thuật toán học có giám sát điển hình, bao gồm các thuật toán phân lớp và hồi quy, một số thuật toán tối ưu bổ trợ. Một số mô hình học máy có giám sát được đề cập bao gồm:

- Hồi quy tuyến tính
- K lân cận
- Bộ phân loại Naive Bayes
- Hạ Gradient
- Thuật toán học Perceptron - PLA
- Hồi quy Logistic
- Hồi quy Softmax
- Máy véc tơ hỗ trợ - SVM

Hồi quy tuyến tính

Quay lại ví dụ đơn giản được nêu trong bài trước: một căn nhà rộng x_1 m^2 có x_2 phòng ngủ và cách trung tâm thành phố x_3 km. có giá là bao nhiêu. Giả sử chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố đó, liệu rằng khi có một căn nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, chúng ta có thể dự đoán được giá của căn nhà đó không? Nếu có thì hàm dự đoán $y = f(\mathbf{x})$ sẽ có dạng như thế nào. Ở đây $\mathbf{x} = [x_1, x_2, x_3]$ là một vector hàng chứa thông tin input,y là một số vô hướng biểu diễn giá nhà. Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào là:

$$y \approx f(\mathbf{x}) = \hat{y}$$

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$$

Phân tích toán học

Trong phương trình nêu trên, nếu chúng ta đặt $\mathbf{w} = [w_0, w_1, w_2, w_3]^T =$ là vector (cột) hệ số cần phải tối ưu và $\bar{\mathbf{x}} = [1, x_1, x_2, x_3]$ (đọc là x bar trong tiếng Anh) là vector (hàng) dữ liệu đầu vào mở rộng, phương trình trên có thể viết gọn lại thành:

$$y \approx \bar{\mathbf{x}}\mathbf{w} = \hat{y}$$

Sai số dự đoán Chúng ta mong muốn rằng sự sai khác e giữa giá trị thực y và $\hat{y}$ (đọc là y hat trong tiếng Anh) là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:

$$\frac{1}{2}e^2 = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \bar{\mathbf{x}}\mathbf{w})^2$$

Hàm mất mát

Điều tương tự xảy ra với tất cả các cặp (input, outcome) $(\mathbf{x}_i, y_i), i = 1, 2, \ldots, N$ với N là số lượng dữ liệu quan sát được. Điều chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm $\mathbf{w}$ để hàm số sau đạt giá trị nhỏ nhất:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \bar{\mathbf{x}}_\mathbf{i}\mathbf{w})^2$$

Hàm số $\mathcal{L}(\mathbf{w})$ được gọi là hàm mất mát (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số $\mathbf{w}$ sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Giá trị của $\mathbf{w}$ làm cho hàm mất mát đạt giá trị nhỏ nhất được gọi là điểm tối ưu (optimal point), ký hiệu:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

Hàm mất mát

rước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát phía trên. Đặt $\mathbf{y} = [y_1; y_2; ...; y_N]$ là một vector cột chứa tất cả các output của training data; $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1; \bar{\mathbf{x}}_2; ...; \bar{\mathbf{x}}_N]$ là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nó là một điểm dữ liệu. Khi đó hàm số mất mát $\mathcal{L}(\mathbf{w})$ được viết dưới dạng ma trận đơn giản hơn:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \bar{\mathbf{x}}_i\mathbf{w})^2 = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2$$

với $\|\mathbf{z}\|_2$ là Euclidean norm (chuẩn Euclid, hay khoảng cách Euclid), nói cách khác $\|\mathbf{z}\|_2^2$ là tổng của bình phương mỗi phần tử của vector $\mathbf{z}$.

Để giải phương trình trên chúng ta lấy đạo hàm theo $\mathbf{w}$ và giải phương trình đạo hàm bằng 0 và lấy nghiệm thu được chính là nghiệm tối ưu.

Bài tập mô phỏng với Python và R

1. Bài tập mô phỏng với Python có thể xem nội dung code trong Bài giảng học máy
2. Nội dung code trong R bấm vào link sau để thực hành: https://rpubs.com/tranquangquy_ictu/794232

Hồi quy đơn biến

Mô hình hồi quy đơn biến là mô hình bao gồm 1 biến phụ thuộc và 1 biến độc lập, công thức chung như sau:

$$Y = \alpha \times X + \beta$$

trong đó α được gọi là hệ số góc của mô hình và β được gọi là hệ số chặn của mô hình. Xét tập dữ liệu có tên women miêu tả về chiều cao và cân nặng của một số cá nhân được đo đạc lại, chúng ta cần tìm một phương trình hồi quy tuyến tính để dự báo biến cân nặng thông qua chiều cao. Tập dữ liệu có mô tả như sau:

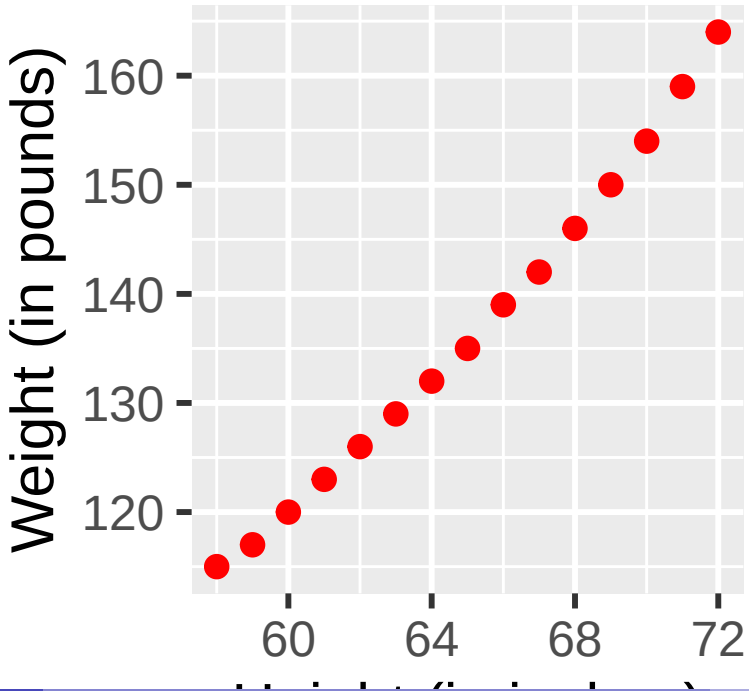
Dữ liệu

```
data("women")
head(women)
```

```
##   height weight
## 1     58    115
## 2     59    117
## 3     60    120
## 4     61    123
## 5     62    126
## 6     63    129
```

Trực quan dữ liệu

```
library(ggplot2)
library(dplyr)
women %>% ggplot(aes(x = height, y = weight)) +
  geom_point( color = "red") +
  xlab("Height (in inches)") + # 1 inch = 2.54 cm
  ylab("Weight (in pounds)") # 1 pound = 0.45 kg
```



Mô hình

```
fit <- lm(weight ~ height, data = women )
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
```

Giá trị dự báo

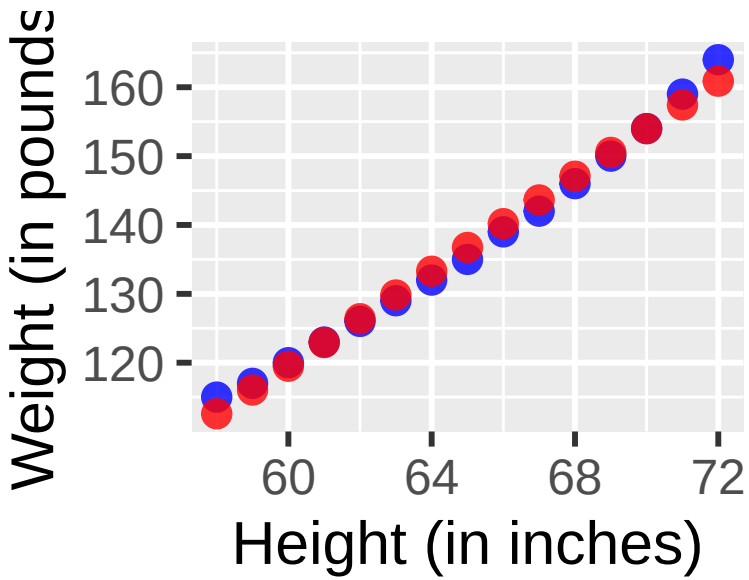
```
library(dplyr)
fitted(fit) %>%
  head()
```

```
##        1        2        3        4        5        6
## 112.5833 116.0333 119.4833 122.9333 126.3833 129.8333
```

```
women$predicton_value <- fitted(fit)
women$residual_value <- residuals(fit)
```

So sánh

```
women %>% ggplot() +
  geom_point(aes(x = height, y = weight),
             col = "blue", size =2, alpha = 0.8) +
  geom_point(aes(x = height, y = predicton_value),
             col = "red", size = 2, alpha = 0.8) +
  xlab("Height (in inches)") + # 1 inch = 2.54 cm
  ylab("Weight (in pounds)") # 1 pound = 0.45 kg
```



Hồi quy đa biến

Khái niệm

Hồi quy đa biến là một phần mở rộng của hồi quy tuyến tính đơn giản. Nó được sử dụng khi chúng ta muốn dự đoán giá trị của một biến dựa trên giá trị của hai hoặc nhiều biến khác. Biến chúng ta muốn dự đoán được gọi là biến phụ thuộc (hoặc đôi khi, biến kết quả, mục tiêu hoặc biến tiêu chí). Các biến chúng ta đang sử dụng để dự đoán giá trị của biến phụ thuộc được gọi là biến độc lập. Hồi quy đa biến cũng cho phép bạn xác định mức độ đóng góp nhiều, ít, không đóng góp... của từng nhân tố vào sự thay đổi của biến phụ thuộc.

Ví dụ: Thu nhập, địa điểm sinh sống và số thành viên trong gia đình ảnh hưởng đến chi tiêu.

- Biến độc lập:Thu nhập, địa điểm, số thành viên
- Biến phụ thuộc: Chi tiêu

Các đại lượng đo sai số trung bình

SSE

SSE là tổng bình phương sai số - Sum Square Error, được tính bằng công thức:

$$SSE = \sum (f_i - y_i)^2$$

RMSE

RMSE - Root Mean Square Error- Căn bậc hai trung bình sai số, tính bằng công thức:

$$RMSE = \sqrt{\frac{\sum_i^n (f_i - y_i)^2}{n}}$$

Hệ số R bình phương

Đây là hệ số thường được sử dụng trong việc đánh giá các mô hình hồi quy với nhau, lựa chọn ra mô hình tốt nhất, tư tưởng của mô hình này được giải thích qua hình sau:

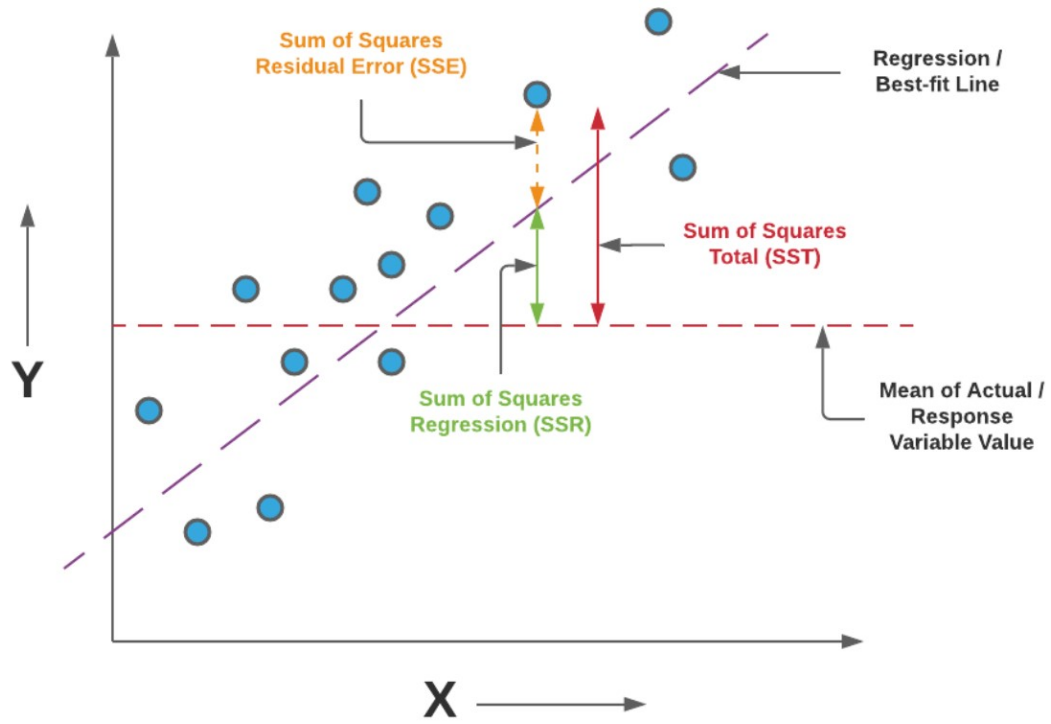


Figure 1: Các sai số bình phương

Hệ số R bình phương

Công thức

Hệ số R^2 này không cung cấp thông tin về độ chính xác của mô hình nhưng cho biết mô hình giải thích được bao nhiêu phần phương sai của biến kết quả trong mẫu, tức là cho chúng ta biết khả năng dự đoán lệch của mô hình với giá trị thực là bao nhiêu.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{\sum(f_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}$$

Ý nghĩa hệ số R bình phương

Ý nghĩa

Giá trị hệ số R^2 có ý nghĩa như sau:

- Nếu giá trị R^2 bằng 0 tức là mô hình hồi quy xây dựng không dự đoán tốt hơn mô hình cơ sở (baseline model - tức là giá trị trung binh $\bar{y}$, dự đoán không cần biến độc lập).
- Nếu giá trị R^2 bằng 1 tức là mô hình dự đoán đúng hoàn toàn chính xác.
- Hệ số R^2 càng cao thì mô hình càng có khả năng dự báo tốt.
- Giá trị R^2 nằm trong khoảng [0,1].

Dạng mô hình hồi quy đa biến

Mô hình hồi quy đa biến dạng tổng quát có dạng:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

Thực hành bài tập hồi quy đa biến theo đường link sau:

Bấm vào đây để xem

Lựa chọn mô hình hồi quy đa biến phù hợp

Một số chú ý

Khi lựa chọn mô hình hồi quy đa biến chúng ta cần chú ý mấy nội dung sau:

- Không nên đưa nhiều biến số vào mô hình khi dữ liệu có ít hoặc đưa biến số không có ý nghĩa dự báo trong mô hình hồi quy.
- Hiện tượng overfitting có thể xảy ra.
- Xảy ra hiện tượng đa cộng tuyến xảy ra khi các biến độc lập có tương quan lớn với nhau

Để tìm hiểu thêm cách chọn mô hình hồi quy đa biến phù hợp, xem ở đây: Hồi quy đa biến