Linear Regression

Lưu Văn Việt

Khoa Toán - Cơ - Tin học
Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia Hà Nội

Ngày 23 tháng 2 năm 2023

Nội dung

Giới thiệu về phân tích hồi quy

Phân tích hồi quy

Phân tích hồi quy thuộc lớp các bài toán học có giám sát của học máy.

Giới thiệu về phân tích hồi quy

Phân tích hồi quy

Phân tích hồi quy thuộc lớp các bài toán học có giám sát của học máy.
Phân tích hồi quy nghiên cứu sự phụ thuộc của một biến phản hồi vào một hay nhiều biến dự báo.

Giới thiệu về phân tích hồi quy

Phân tích hồi quy

Phân tích hồi quy thuộc lớp các bài toán học có giám sát của học máy.
Phân tích hồi quy nghiên cứu sự phụ thuộc của một biến phản hồi vào một hay nhiều biến dự báo.

- y: Biến phản hồi - Luôn là biến liên tục
- $(x_1, x_2, ..., x_N) \subset \mathbb{R}^d, x_i = (x_i^1, x_i^2, .., x_i^d) \in \mathbb{R}^d$

Giới thiệu về phân tích hồi quy

Phân tích hồi quy

Phân tích hồi quy thuộc lớp các bài toán học có giám sát của học máy.
Phân tích hồi quy nghiên cứu sự phụ thuộc của một biến phản hồi vào một hay nhiều biến dự báo.

- y: Biến phản hồi - Luôn là biến liên tục
- $(x_1, x_2, ..., x_N) \subset \mathbb{R}^d, x_i = (x_i^1, x_i^2, .., x_i^d) \in \mathbb{R}^d$

Phân loại phân tích hồi quy

Phân loại

Có 2 dạng phân tích hồi quy:

Phân loại phân tích hồi quy

Phân loại

Có 2 dạng phân tích hồi quy:

- Phân tích hồi quy tuyến tính
- Phân tích hồi quy phi tuyến

Phân loại phân tích hồi quy

Phân loại

Có 2 dạng phân tích hồi quy:

- Phân tích hồi quy tuyến tính
- Phân tích hồi quy phi tuyến

Phân loại theo biến phản hồi y:

- Phân tích hồi quy đơn: $y \in \mathbb{R}$
- Phân tích hồi quy bội: $y \in \mathbb{R}^n, \quad n \geq 1$

Phân loại phân tích hồi quy

Phân loại

Có 2 dạng phân tích hồi quy:

- Phân tích hồi quy tuyến tính
- Phân tích hồi quy phi tuyến

Phân loại theo biến phản hồi y:

- Phân tích hồi quy đơn: $y \in \mathbb{R}$
- Phân tích hồi quy bội: $y \in \mathbb{R}^n, \quad n \geq 1$

Phân loại theo biến dự báo x:

- Phân tích hồi quy đơn biến: $x \in \mathbb{R}$
- Phân tích hồi quy nhiều biến: $x \in \mathbb{R}^d, \quad d \geq 1$

Phân loại phân tích hồi quy

Phân loại

Có 2 dạng phân tích hồi quy:

- Phân tích hồi quy tuyến tính
- Phân tích hồi quy phi tuyến

Phân loại theo biến phản hồi y:

- Phân tích hồi quy đơn: $y \in \mathbb{R}$
- Phân tích hồi quy bội: $y \in \mathbb{R}^n, \quad n \geq 1$

Phân loại theo biến dự báo x:

- Phân tích hồi quy đơn biến: $x \in \mathbb{R}$
- Phân tích hồi quy nhiều biến: $x \in \mathbb{R}^d, \quad d \geq 1$

Phân tích hồi quy tuyến tính - Linear Regression

Linear Regression

Trong phân tích hồi quy tuyến tính, hàm dự báo $h_\theta(x)$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + ... + \theta_d x_d$$

Linear Regression

Trong phân tích hồi quy tuyến tính, hàm dự báo $h_\theta(x)$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + ... + \theta_d x_d$$

Nếu bổ sung thêm đặc trưng cố định $x0 \equiv 1$ thì ta có thể biểu diễn h_θ dưới dạng:

Phân tích hồi quy tuyến tính - Linear Regression

Linear Regression

Trong phân tích hồi quy tuyến tính, hàm dự báo $h_\theta(x)$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + ... + \theta_d x_d$$

Nếu bổ sung thêm đặc trưng cố định $x0 \equiv 1$ thì ta có thể biểu diễn h_θ dưới dạng:

$$h_\theta(x) = \sum_{i=0}^{d} \theta_i x_i$$

Phân tích hồi quy tuyến tính - Linear Regression

Linear Regression

Trong phân tích hồi quy tuyến tính, hàm dự báo $h_\theta(x)$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + ... + \theta_d x_d$$

Nếu bổ sung thêm đặc trưng cố định $x0 \equiv 1$ thì ta có thể biểu diễn h_θ dưới dạng:

$$h_\theta(x) = \sum_{i=0}^{d} \theta_i x_i$$

Phân tích hồi quy tuyến tính - Linear Regression

Linear Regression

Trong phân tích hồi quy tuyến tính, hàm dự báo $h_\theta(x)$ được xấp xỉ bởi một hàm tuyến tính của x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + ... + \theta_d x_d$$

Nếu bổ sung thêm đặc trưng cố định $x0 \equiv 1$ thì ta có thể biểu diễn h_θ dưới dạng:

$$h_\theta(x) = \sum_{i=0}^{d} \theta_i x_i$$

Linear Regression

Với $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_d \end{pmatrix}, x = \begin{pmatrix} 1 \\ x_1 \\ \dots \\ x_d \end{pmatrix}$, ta có biểu diễn của $h_\theta(x)$ như sau:

$$h_\theta(x) = \theta^T x$$

Linear Regression

Với $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_d \end{pmatrix}, x = \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_d \end{pmatrix}$, ta có biểu diễn của $h_\theta(x)$ như sau:

$$h_\theta(x) = \theta^T x$$

Để ước lượng tham số θ, ta cực tiểu hóa sai số của mô hình trên tập dữ liệu huấn luyện: training_set $= \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$

Linear Regression

Với $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_d \end{pmatrix}, x = \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_d \end{pmatrix}$, ta có biểu diễn của $h_\theta(x)$ như sau:

$$h_\theta(x) = \theta^T x$$

Để ước lượng tham số θ, ta cực tiểu hóa sai số của mô hình trên tập dữ liệu huấn luyện: training_set $= \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$

$$J(\theta) = \sum_{i=1}^{N} (h_\theta(x_i) - y_i)^2 \rightarrow \min_\theta J(\theta)$$

Linear Regression

Với $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ ... \\ \theta_d \end{pmatrix}, x = \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_d \end{pmatrix}$, ta có biểu diễn của $h_\theta(x)$ như sau:

$$h_\theta(x) = \theta^T x$$

Để ước lượng tham số θ, ta cực tiểu hóa sai số của mô hình trên tập dữ liệu huấn luyện: training_set $= \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$

$$J(\theta) = \sum_{i=1}^{N} (h_\theta(x_i) - y_i)^2 \rightarrow \min_\theta J(\theta)$$

Linear Regression

Linear Regression

Với $x_i \in \mathbb{R}^d$, xét ma trận:

$$X \in \mathcal{M}_{N\times(d+1)} = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

Linear Regression

Với $x_i \in \mathbb{R}^d$, xét ma trận:

$$X \in \mathcal{M}_{N\times(d+1)} = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

Khi đó: $J(\theta) = ||X\theta - y||_2^2$

$$\begin{aligned} &\Rightarrow \quad J(\theta) \to \min \Leftrightarrow ||X\theta - y||_2 \to \min \\ &\Leftrightarrow \quad \theta = \hat{\theta}, \quad X^T X\hat{\theta} = X^T y \\ &\Leftrightarrow \quad \hat{\theta} = (X^T X)^{-1} X^T y \end{aligned}$$

Linear Regression

Với $x_i \in \mathbb{R}^d$, xét ma trận:

$$X \in \mathcal{M}_{N\times(d+1)} = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

Khi đó: $J(\theta) = ||X\theta - y||_2^2$

$$\begin{aligned} &\Rightarrow \quad J(\theta) \to \min \Leftrightarrow ||X\theta - y||_2 \to \min \\ &\Leftrightarrow \quad \theta = \hat{\theta}, \quad X^T X\hat{\theta} = X^T y \\ &\Leftrightarrow \quad \hat{\theta} = (X^T X)^{-1} X^T y \end{aligned}$$

Phương trình $\hat{\theta} = (X^T X)^{-1} X^T y$ được gọi là phương trình chuẩn.

Linear Regression

Với $x_i \in \mathbb{R}^d$, xét ma trận:

$$X \in \mathcal{M}_{N\times(d+1)} = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_N^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

Khi đó: $J(\theta) = ||X\theta - y||_2^2$

$$\begin{aligned} &\Rightarrow \quad J(\theta) \to \min \Leftrightarrow ||X\theta - y||_2 \to \min \\ &\Leftrightarrow \quad \theta = \hat{\theta}, \quad X^T X\hat{\theta} = X^T y \\ &\Leftrightarrow \quad \hat{\theta} = (X^T X)^{-1} X^T y \end{aligned}$$

Phương trình $\hat{\theta} = (X^T X)^{-1} X^T y$ được gọi là phương trình chuẩn.

Xác định $\hat{\theta}$ bằng phân tích QR

Xác định $\hat{\theta}$ bằng phân tích QR

Trong thực tế, chúng ta ít khi sử dụng trực tiếp phương trình chuẩn để xác định $\hat{\theta}$ do phép lấy nghịch đảo ma trận $(X^T X)^{-1}$ có thể dẫn tới các sai số làm tròn lớn trong quá trình tính toán.

Chúng ta thường sử dụng phương pháp QR để xác định nghiệm $\hat{\theta}$ của phương trình:

$$X^T X\theta = X^T y$$

Xác định $\hat{\theta}$ bằng phân tích QR

Xác định $\hat{\theta}$ bằng phân tích QR

Ta tìm cách phân tích ma trận $X = QR$, trong đó Q là ma trận trực giao còn R là ma trận tam giác trên.

Xác định $\hat{\theta}$ bằng phân tích QR

Ta tìm cách phân tích ma trận $X = QR$, trong đó Q là ma trận trực giao còn R là ma trận tam giác trên.

Phân tích QR rút gọn:

$$X = \hat{Q}\hat{R}$$

trong đó $\hat{Q}$ là ma trận vuông gồm $(d+1)$ cột đầu của Q còn $\hat{R}$ là ma trận vuông gồm $(d+1)$ hàng đầu của R.

Xác định $\hat{\theta}$ bằng phân tích QR

Ta tìm cách phân tích ma trận $X = QR$, trong đó Q là ma trận trực giao còn R là ma trận tam giác trên.

Phân tích QR rút gọn:

$$X = \hat{Q}\hat{R}$$

trong đó $\hat{Q}$ là ma trận vuông gồm $(d+1)$ cột đầu của Q còn $\hat{R}$ là ma trận vuông gồm $(d+1)$ hàng đầu của R.

Khi đó, bài toán ban đầu của chúng ta trở thành tìm θ là nghiệm của hệ:

$$\hat{R}\theta = \hat{Q}^T b$$

Xác định $\hat{\theta}$ bằng phân tích QR

Ta tìm cách phân tích ma trận $X = QR$, trong đó Q là ma trận trực giao còn R là ma trận tam giác trên.
Phân tích QR rút gọn:

$$X = \hat{Q}\hat{R}$$

trong đó $\hat{Q}$ là ma trận vuông gồm $(d+1)$ cột đầu của Q còn $\hat{R}$ là ma trận vuông gồm $(d+1)$ hàng đầu của R.
Khi đó, bài toán ban đầu của chúng ta trở thành tìm θ là nghiệm của hệ:

$$\hat{R}\theta = \hat{Q}^T b$$

Do $\hat{R}$ là ma ma trận tam giác trên nên nghiệm $\hat{\theta}$ có thể dễ dàng xác định được bằng quá trình thế ngược.

Xác định $\hat{\theta}$ bằng phân tích QR

Ta tìm cách phân tích ma trận $X = QR$, trong đó Q là ma trận trực giao còn R là ma trận tam giác trên.
Phân tích QR rút gọn:

$$X = \hat{Q}\hat{R}$$

trong đó $\hat{Q}$ là ma trận vuông gồm $(d+1)$ cột đầu của Q còn $\hat{R}$ là ma trận vuông gồm $(d+1)$ hàng đầu của R.
Khi đó, bài toán ban đầu của chúng ta trở thành tìm θ là nghiệm của hệ:

$$\hat{R}\theta = \hat{Q}^T b$$

Do $\hat{R}$ là ma ma trận tam giác trên nên nghiệm $\hat{\theta}$ có thể dễ dàng xác định được bằng quá trình thế ngược.

Tìm phân tích QR rút gọn bằng thuật toán trực giao hóa Gram-Schmidt

Phân tích hồi quy tuyến tính - Linear Regression

Tìm phân tích QR rút gọn bằng thuật toán trực giao hóa Gram-Schmidt

- Bước 1:
  - $+ r_{11} = \sqrt{X_1^T X_1}$
  - $+ q_1 = \frac{1}{r_{11}} X_1$
- Bước $k = 2 \to (d+1)$:
  - $+ v = X_k$
  - $+$ *Với* $i = 1 \to k-1$
    - $+) \ r_{ik} = q_i^T v$
    - $+) \ v = v - r_{ik} q_i$
  - $+ r_{kk} = \sqrt{v^T v}$
  - $+ q_k = \frac{v}{r_{kk}}$

Phân tích hồi quy tuyến tính - Linear Regression

Tìm phân tích QR rút gọn bằng thuật toán trực giao hóa Gram-Schmidt

- Bước 1:
  - $+ r_{11} = \sqrt{X_1^T X_1}$
  - $+ q_1 = \frac{1}{r_{11}} X_1$
- Bước $k = 2 \to (d+1)$:
  - $+ v = X_k$
  - $+$ *Với* $i = 1 \to k-1$
    - $+) \ r_{ik} = q_i^T v$
    - $+) \ v = v - r_{ik} q_i$
  - $+ r_{kk} = \sqrt{v^T v}$
  - $+ q_k = \frac{v}{r_{kk}}$

Phân tích hồi quy tuyến tính bằng thư viện sklearn

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

+ Huấn luyện mô hình:

```
regr = LinearRegression(fit_intercept= False)
regr.fit(Xbar, y_train)
```

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

+ Huấn luyện mô hình:

```
regr = LinearRegression(fit_intercept= False)
regr.fit(Xbar, y_train)
```

+ Hiển thị hệ số của mô hình hồi quy:

```
regr.coef_
```

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

+ Huấn luyện mô hình:

```
regr = LinearRegression(fit_intercept= False)
regr.fit(Xbar, y_train)
```

+ Hiển thị hệ số của mô hình hồi quy:

```
regr.coef_
```

+ Chạy mô hình trên tập test:

```
y_pred = regr.predict(X_test)
```

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

+ Huấn luyện mô hình:

```
regr = LinearRegression(fit_intercept= False)
regr.fit(Xbar, y_train)
```

+ Hiển thị hệ số của mô hình hồi quy:

```
regr.coef_
```

+ Chạy mô hình trên tập test:

```
y_pred = regr.predict(X_test)
```

+ Đánh giá mô hình bằng MSE:

```
mean_square_error(y_test, y_pred)
```

Phân tích hồi quy tuyến tính - Linear Regression

Phân tích hồi quy tuyến tính bằng thư viện sklearn

+ Sử dụng thư viện:

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_metrics import mean_square_error
```

+ Huấn luyện mô hình:

```
regr = LinearRegression(fit_intercept= False)
regr.fit(Xbar, y_train)
```

+ Hiển thị hệ số của mô hình hồi quy:

```
regr.coef_
```

+ Chạy mô hình trên tập test:

```
y_pred = regr.predict(X_test)
```

+ Đánh giá mô hình bằng MSE:

```
mean_square_error(y_test, y_pred)
```

Ví dụ

Ví dụ

Trong tệp dữ liệu SAT_GPA.csv có 84 mẫu dữ liệu điểm thi của các sinh viên, mẫu có 02 trường dữ liệu, trong cột thứ nhất chứa trường điểm SAT (Reading + Mathematic + Writing) của các kỳ thi trong bậc phổ thông; cột thứ hai chứa điểm trung bình GPA của sinh viên tương ứng ở bậc học đại học/cao đẳng. Chúng ta xây dựng một mô hình hồi quy tuyến tính để mô tả sự phụ thuộc của điểm GPA ở bậc đại học/cao đẳng vào điểm SAT của mỗi sinh viên ở bậc phổ thông. Sử dụng 60 bộ dữ liệu đầu tiên để huấn luyện và 24 bộ dữ liệu còn lại để đánh giá mô hình.

Áp dụng phương pháp hồi quy ta thu được mô hình:

$$GPA = 0.88948508 + 0.0012857 \times SAT$$

Trung bình bình phương sai số của mô hình là:

$$MSE = 0.06994526432657806$$