ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KỲ MÔN HỌC

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

ĐỒ ÁN: DATA FITTING

Giáo viên hướng dẫn	Thầy Nguyễn Hữu Toàn
	Cô Võ Nam Thục Đoan
Lớp	21_2
Sinh viên thực hiện	21120150 – Nguyễn Song Toàn

Tp. Hồ Chí Minh, tháng 06 năm 2023

Lời cảm ơn
Em xin cảm ơn thầy Nguyễn Hữu Toàn đã cung cấp cho chúng em những kiến thức bổ ích, có ý nghĩa thực tiễn để em có thể thực hiện đồ án này. Em cũng xin cảm ơn cô Võ Nam Thục Đoan đã cung cấp những bài thực hành để em vận dụng được những lý thuyết đã học và có những kỹ thuật lập trình phục vụ cho việc thực hiện các đồ án.

Mục lục

Danh mục hình

A. Thông tin đồ án

1. Đề bài

- Sử dụng bài toán data fitting trong xây dựng mô hình đánh giá lương nhân viên từ các yếu tố tác động theo dữ liệu được cung cấp.
- Sử dụng bài toán data fitting trong việc xây dựng mô hình đánh giá giá nhà từ các yếu tố tác động từ dữ liệu được cung cấp.

2. Đánh giá mức độ hoàn thành

Nội dung	Mức độ hoàn thành
Xây dựng mô hình đánh giá lương nhân viên	100%
Xây dựng mô hình đánh giá giá nhà	100%

B. Cơ sở lý thuyết

1. Bài toán bình phương tối tiểu (Least square)

Xét ma trận **A** có kích thước $m \times n$ ($m > n$) và vector **b** có kích thước m. Bình phương tối tiểu $\hat{\boldsymbol{r}}$ của **A** và **b** trên giá trị x được tính như sau:

$$\hat{r} = \|Ax - b\|^2$$
$$\hat{x} = (A^T A)^{-1} A^T b$$

2. Mô hình hồi quy tuyến tính (Linear regression model)

- Nghiệm $\hat{x}$ của phương trình hồi quy tuyến tính $\boldsymbol{Ax = b}$ được tính như sau:

$$\hat{x} = (A^T A)^{-1} A^T b$$

- Cách xác định các tham số của mô hình hồi quy tuyến tính:
  Cho y là biến phụ thuộc, $x_1, x_2, x_3, \ldots, x_n$ là biến độc lập.
  Tìm 1 hàm số biểu diễn $y = f(x_1, x_2, \ldots, x_n)$
  Gọi $y = b_1x_1 + b_2x_2 + \ldots + b_nx_n$
  Ta có: $RSS(\theta) = \|\hat{y}_i - y_i\|^2$
  Trong đó:
  - $\hat{y}$: giá trị được tính bằng cách thay biến độc lập vào hàm số
  - y: giá trị thực biến phụ thuộc

  $RSS(\theta) = \|A\theta - Y\|^2$

$$A = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{21} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{nn} \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

  Tìm θ sao cho $\|A\theta - Y\|^2$ min

Khi đó: $\theta = (A^T A)^{-1}(A^T Y)$

Từ đó, xác định các tham số của mô hình hồi quy tuyến tính.

- Cách tính chuẩn vector phần dư:

  Sau khi tìm được các tham số thì tính:

$$\widehat{Y} = \begin{pmatrix} Y(x_1) \\ Y(x_2) \\ \vdots \\ Y(x_n) \end{pmatrix} \Rightarrow r = \widehat{Y} - Y = \begin{pmatrix} Y(x_1) - y_1 \\ Y(x_2) - y_2 \\ \vdots \\ Y(x_n) - y_n \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

$$\Rightarrow \|r\| = \sqrt{r_1^2 + r_2^2 + \ldots + r_n^2}$$

C. Mô hình sử dụng ở câu a và câu b

1. Câu a - mô hình đánh giá giá nhà từ các yếu tố tác động

$$\ln(Y) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

- x_1 đại diện cho đặc trưng sqft / 100
- x_2 đại diện cho đặc trưng Bedrooms
- x_3 đại diện cho đặc trưng Baths
- x_4 đại diện cho đặc trưng Age

2. Câu b - mô hình đánh giá lương nhân viên từ các yếu tố tác động

$$\ln(Y) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

- x_1 đại diện cho đặc trưng educ
- x_2 đại diện cho đặc trưng exper
- x_3 đại diện cho đặc trưng hrswk

D. Mô tả ý nghĩa của các hàm

- Hàm **create_identity_matrix(n)**: tạo ra ma trận đơn vị I_n.
- Hàm **multiply_matrix_vector(matrix, vector)**: trả ra kết quả của nhân ma trận với vector.
- Hàm **multiply_two_matrixes(A, B)**: trả ra kết quả của nhân 2 ma trận.
- Hàm **transpose_matrix(A)**: trả ra ma trận chuyển vị của ma trận A.
- Hàm **swap_two_rows(matrix, r1, r2)**: hoán vị 2 dòng của ma trận với tham số là ma trận matrix, chỉ số của 2 dòng r1 và r2.
- Hàm **inverse(A)**: trả ra ma trận nghịch đảo của ma trận A (nếu có).
- Hàm **subtract_two_vectors(A, B)**: trả ra kết quả của trừ 2 vector A và B.
- Hàm **create_y_A(data)**: tạo ra vector y và ma trận A dựa trên bộ dữ liệu data được đọc từ file dùng cho câu a.

- Hàm này sử dụng vòng lặp để tạo ra vector y mang giá trị *ln(price)* và ma trận A gồm cột đầu tiên chứa toàn số 1 và các cột tiếp theo lần lượt là cột *sqft/100, Bedrooms, Baths, Age*, các cột này là giá trị của các cột trong dataframe mà đọc vào từ file br2.csv.

- Hàm **vectorTheta(A, y)**: xây dựng vector các tham số của mô hình dựa trên cơ sở lý thuyết đã học.
  - Hàm này sử dụng các hàm hỗ trợ tính toán ma trận, xây dựng ma trận nghịch đảo như: **multiply_two_matrixes(A, B), multiply_matrix_vector(matrix, vector), inverse(A)** để thực hiện tìm ra vector tham số theo công thức như sau:

$$\theta = (A^T A)^{-1}(A^T Y)$$

- Hàm **getModel_a(Theta, x)**: trả về chuỗi là hàm số của mô hình khi có danh sách biến x và vector tham số Theta.
  - Hàm này sử dụng vòng lặp và xử lý chuỗi để có được chuỗi theo định dạng như sau: "Y* = lnY = $\theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \theta_3 * x_3$" với θ_i là các giá trị tham số cụ thể.
- Hàm **exportModel_a(A, y, n)**: xây dựng danh sách biến dựa trên số lượng đặc trưng n, xây dựng vector Theta từ ma trận A và vector y và trả ra hàm số của mô hình khi có được danh sách biến và danh sách tham số Theta.
- Hàm **vectorYhat(A, Theta)**: tạo ra vector $\hat{Y}$ từ ma trận A và vector Theta để phục vụ cho việc tính sai số của mô hình (tính chuẩn).
- Hàm **get_error(y, A)**: trả ra chuẩn vector phần dư của mô hình.
- Hàm **create_y_A_b(data)**: tạo ma trận A và vector y cho câu b.
  - Hàm này sử dụng vòng lặp để tạo ra vector y mang giá trị *ln(wage)* và ma trận A gồm cột đầu tiên chứa toàn số 1 và các cột tiếp theo lần lượt là cột *educ, exper, hrswk*, các cột này là giá trị của các cột trong dataframe mà đọc vào từ file cps4_small.csv.
- Trong đồ án này, em sử dụng thư viện *pandas* để đọc dữ liệu từ file .csv và dùng thư viện *math* để tính log.

E. Kết quả đồ án

- Mô hình đánh giá giá nhà từ các yếu tố tác động:

```
Mô hình đánh giá giá nhà:
Y* = lnY = 10.91896  + 0.03308 * x1 - 0.05895 * x2 + 0.21457 * x3 - 0.00660 * x4
Chuẩn vector phần dư là: 9.12
```

Hình 1 - Mô hình đánh giá giá nhà

- **Mô hình đánh giá lương nhân viên từ các yếu tố tác động:**

```
Mô hình đánh giá lương nhân viên:
Y* = lnY = 1.10054  + 0.09031 * x1 + 0.00578 * x2 + 0.00894 * x3
Chuẩn vector phần dư là: 16.21
```

Hình 2 - Mô hình đánh giá lương nhân viên

Tài liệu tham khảo

[1] Slide Toán ứng dụng và thống kê, trường đại học Khoa học tự nhiên, ĐHQG-HCM

[2] Tài liệu hướng dẫn lab Linear Regression, trường đại học Khoa học tự nhiên, ĐHQG-HCM