# EVALUATING REGRESSION MODELS PERFORMANCE

1. TS. Nguyễn Tấn Trần Minh Khang
2. ThS. Võ Duy Nguyên
3. Cao học. Nguyễn Hoàn Mỹ
4. Tình nguyện viên. Lê Ngọc Huy
5. Tình nguyện viên. Cao Bá Kiệt

# Đánh giá mô hình hồi quy

– Có nhiều cách để đánh giá một mô hình hồi quy:

+ Cách 01: Trực quan hóa kết quả mô hình và dữ liệu.

+ Cách 02: Hàm lỗi Squared Sum (SSE).

+ Cách 03: Hàm lỗi Root Mean Squared (RMSE).

+ Cách 04: Hàm đánh giá $R^2$.

+ Cách 05: Hàm đánh giá $R^2_{adjusted}$.

# HÀM LỖI SQUARED SUM (SSE)

# Hàm lỗi Squared Sum (SSE)

– Hàm lỗi *Squared Sum* được định nghĩa như sau:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

– Trong đó:

+ $n$ là số lượng điểm dữ liệu trong tập dữ liệu.
+ $y_i$ là kết quả thực của điểm dữ liệu thứ $i$.
+ $\hat{y}_i$ là kết quả dự đoán bởi mô hình của điểm dữ liệu thứ $i$.

# Hàm lỗi Squared Sum (SSE)

- Hàm lỗi *Squared Sum* có giá trị trong khoảng $[0, +\infty)$:
  - + $SSE = 0$: Mô hình chính xác tuyệt đối.
  - + Trên cùng một tập dữ liệu, $SSE$ càng nhỏ, độ chính xác của mô hình càng cao.

# Hàm lỗi Squared Sum (SSE)

– Ví dụ trên 5 điểm dữ liệu đầu tiên của tập dữ liệu Position Salaries:

| Level | Salary |
|---|---|
| 1 | 45,000 |
| 2 | 50,000 |
| 3 | 60,000 |
| 4 | 80,000 |
| 5 | 110,000 |

| Predicted Salary |
|---|
| 40,000 |
| 51,000 |
| 60,000 |
| 83,000 |
| 111,500 |

# Hàm lỗi Squared Sum (SSE)

- Ví dụ trên 5 điểm dữ liệu đầu tiên của tập dữ liệu Position Salaries:
  - + $SSE = (45{,}000 - 40{,}000)^2 + (50{,}000 - 51{,}000)^2 + (60{,}000 - 60{,}000)^2 + (80{,}000 - 83{,}000)^2 + (110{,}000 - 111{,}500)^2.$
  - + $SSE = 5000^2 + 1000^2 + 0^2 + 3000^2 + 1500^2 = 37{,}250{,}000.$

# Hàm lỗi Squared Sum (SSE)

- Nhược điểm:
  - + Không cùng đơn vị với kết quả đầu ra.
  - + Giá trị tăng khi số lượng điểm dữ liệu tăng lên.
- Do đó, hàm lỗi Squared Sum sẽ khó khăn trong việc:
  - + So sánh với kết quả đầu ra.
  - + So sánh kết quả mô hình trên hai tập dữ liệu khác nhau (chẳng hạn tập training và tập test).

# Hàm lỗi Squared Sum (SSE)

- Giới thiệu lại tập dữ liệu Salary Data.
- Mỗi điểm dữ liệu mô tả mức lương của một người khi biết số năm kinh nghiệm của họ.

| STT | Year Experience | Salary |
|---|---|---|
| 1 | 1.1 | 39,343 |
| 2 | 1.3 | 46,205 |
| 3 | 1.5 | 37,731 |
| 4 | 2 | 43,525 |
| 5 | 2.2 | 39,891 |
| 6 | 2.9 | 56,642 |
| 7 | 3 | 60,150 |

# Hàm lỗi Squared Sum (SSE)

- Giới thiệu lại tập dữ liệu Salary Data.
- Mỗi điểm dữ liệu mô tả mức lương của một người khi biết số năm kinh nghiệm của họ.

# Hàm lỗi Squared Sum (SSE)

- Ví dụ: Huấn luyện mô hình Linear Regression và Decision Tree Regression trên:
  + Tập training (chiếm 80% điểm dữ liệu) của bộ dữ liệu Salary Data.

Tính giá trị hàm lỗi SSE của 2 mô hình trên tập training và tập test của bộ dữ liệu này.

# Hàm lỗi Squared Sum (SSE)

| | Linear Regression | Decision Tree |
|---|---|---|
| Salary Data (Training set) | $8.67e+08$ | $3.88e+08$ |
| Salary Data (Test set) | $7.69e+07$ | $1.71e+08$ |

# HÀM LỖI ROOT MEAN SQUARED (RMSE)

# Hàm lỗi Root Mean Squared (RMSE)

– Hàm lỗi Root Mean Squared (RMSE) được định nghĩa như sau:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{SSE}{n}}$$

– Trong đó:

+ $n$ là số lượng điểm dữ liệu trong tập dữ liệu.
+ $y_i$ là kết quả thực của điểm dữ liệu thứ $i$.
+ $\hat{y}_i$ là kết quả dự đoán bởi mô hình của điểm dữ liệu thứ $i$.

# Hàm lỗi Root Mean Squared (RMSE)

- $RMSE$ được xem như là độ sai lệch trung bình giữa đầu ra dự đoán với đầu ra thực.
- Hàm lỗi *Root Mean Squared* có giá trị trong khoảng $[0, +\infty)$:
  - + $RMSE = 0$: Mô hình chính xác tuyệt đối.
  - + $RMSE$ càng nhỏ, độ chính xác của mô hình càng cao.

# Hàm lỗi Root Mean Squared (RMSE)

– Ví dụ trên 5 điểm dữ liệu đầu tiên của tập dữ liệu Position Salaries:

| Level | Salary |
|---|---|
| 1 | 45,000 |
| 2 | 50,000 |
| 3 | 60,000 |
| 4 | 80,000 |
| 5 | 110,000 |

| Predicted Salary |
|---|
| 40,000 |
| 51,000 |
| 60,000 |
| 83,000 |
| 111,500 |

# Hàm lỗi Root Mean Squared (RMSE)

- Ví dụ trên 5 điểm dữ liệu đầu tiên của tập dữ liệu Position Salaries:
  - + $SSE = 37{,}250{,}000.$
  - + $RMSE = \sqrt{\frac{SSE}{5}} \approx 2729.469.$

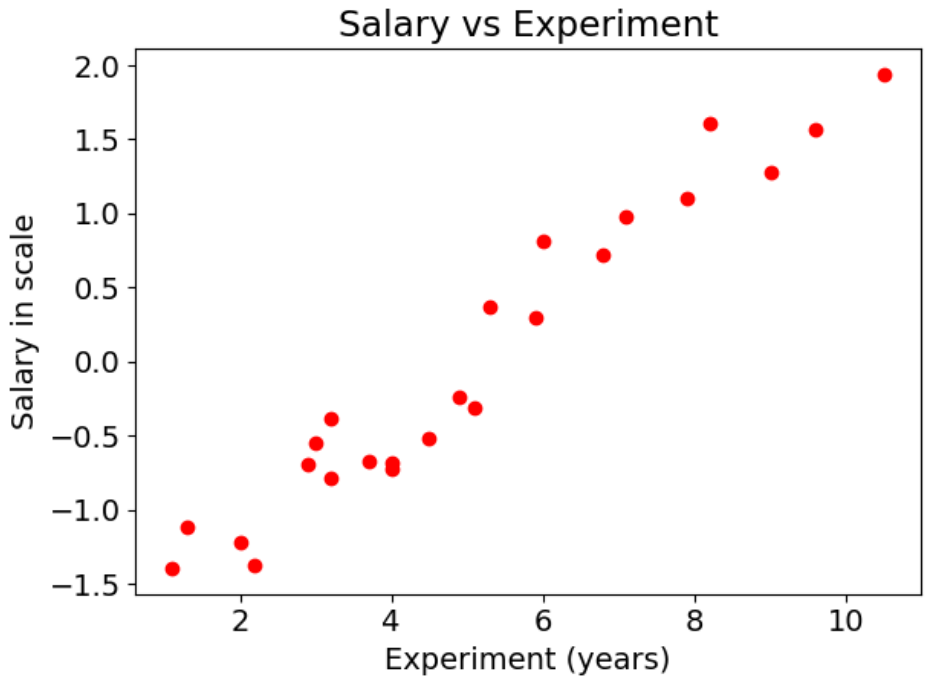
| Level | Salary |
|---|---|
| 1 | 45,000 |
| 2 | 50,000 |
| 3 | 60,000 |
| 4 | 80,000 |
| 5 | 110,000 |

# Hàm lỗi Root Mean Squared (RMSE)

- Hàm lỗi *Root Mean Squared* thường được sử dụng nhiều vì khắc phục được các nhược điểm của hàm lỗi *Squared Sum:*
  - + Không phụ thuộc vào số lượng điểm dữ liệu.
  - + Cùng đơn vị với kết quả đầu ra.
- Nhược điểm:
  - + Phụ thuộc vào miền giá trị đầu ra (outcome) của dữ liệu.

# Hàm lỗi Root Mean Squared (RMSE)

– Giới thiệu tập dữ liệu Salary Data với đầu ra (outcome) đã được chuẩn hóa (nomalized hoặc scale).



# Hàm lỗi Root Mean Squared (RMSE)

- Ví dụ huấn luyện 2 mô hình Linear Regression và Decision Tree Regression trên:
  + Tập training của bộ dữ liệu Salary Data.
  + Tập training với đầu ra đã được chuẩn hóa của bộ dữ liệu Salary Data.

Tính giá trị hàm lỗi RMSE của 2 mô hình trên các tập dữ liệu tương ứng.

# Hàm lỗi Root Mean Squared (RMSE)

| | Linear Regression | Decision Tree |
|---|---|---|
| Salary Data (Training set) | 6,012 | 3,580 |
| Salary Data (Test set) | 4,025 | 5,350 |
| Salary Data (Training set + Scale Outcome) | 0.24 | 0.16 |

# MÔ HÌNH ĐƯỜNG CƠ SỞ - BASELINE

# Mô hình Đường cơ sở – BaseLine

- Trước khi đi vào các cách đánh giá mô hình tiếp theo, ta sẽ giới thiệu về một mô hình đơn giản – gọi là mô hình đường cơ sở (BaseLine).
- Mô hình này chỉ dự đoán một kết quả đầu ra duy nhất đó cho mọi điểm dữ liệu đầu vào, đó là giá trị trung bình của tất cả các kết quả đầu ra trong tập dữ liệu.

# Mô hình Đường cơ sở – BaseLine

– Ví dụ trên 5 điểm dữ liệu đầu tiên của tập dữ liệu Position Salaries:

| Level | Salary |
|---|---|
| 1 | 45,000 |
| 2 | 50,000 |
| 3 | 60,000 |
| 4 | 80,000 |
| 5 | 110,000 |

# Mô hình Đường cơ sở – BaseLine

– Mô hình đường cơ sở của tập dữ liệu này sẽ trả về một kết quả (Salary) duy nhất cho mọi giá trị đầu vào (Level) là:

$$y_{mean} = \frac{45{,}000 + 50{,}000 + 60{,}000 + 80{,}000 + 110{,}000}{5} = 69{,}500$$

# HÀM ĐÁNH GIÁ R SQUARED

# Hàm đánh giá R Squared

– Hàm đánh giá *R Squared* (ký hiệu là $R^2$) được định nghĩa là:

$$R^2 = 1 - \frac{SSE_{model}}{SSE_{baseline}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - y_{mean})^2}$$

– Trong đó:

+ $SSE_{model}$ là giá trị hàm lỗi *Squared Sum* của tập dữ liệu khi đánh giá trên mô hình đang xét.

+ $SSE_{baseline}$ là giá trị hàm lỗi *Squared Sum* của tập dữ liệu khi đánh giá trên mô hình đường cơ sở.

# Hàm đánh giá R Squared

- Giá trị của hệ số $R^2$ luôn nằm trong đoạn $(-\infty, 1]$:
  - + Nếu $R^2 < 0$: Mô hình tệ hơn mô hình đường cơ sở.
  - + Nếu $R^2 = 0$: Mô hình giống như mô hình cơ sở.
  - + Nếu $R^2 = 1$: Mô hình chính xác tuyệt đối.
- $R^2$ càng lớn (càng gần 1) thì độ chính xác của mô hình với tập dữ liệu đang xét càng cao.
- Một mô hình được xem là tốt nếu $R^2 > 0.8$.

# Hàm đánh giá R Squared

- Nhược điểm: Khi sử dụng hàm đánh giá R Squared để so sánh hai mô hình với số lượng đặc trưng đầu vào khác nhau, thì mô hình với số lượng đặc trưng đầu vào lớn hơn (gần như luôn luôn) cho giá trị R Squared lớn hơn.
  - Vì vậy, hàm đánh giá *R squared* sẽ (gần như luôn luôn) nói rằng, mô hình nhận nhiều đặc trưng đầu vào hơn là tốt hơn, cho dù có một số đặc trưng không tương quan với kết quả đầu ra (chẳng hạn đặc trưng *tên* không tương quan với kết quả *mức lương*).
- Khi dữ liệu quá ít, giá trị của hàm R Squared sẽ không ổn định và không đáng tin cậy.

# Hàm đánh giá R Squared

- Ví dụ 01: Huấn luyện mô hình Linear Regression và Decision Tree Regression trên:
  + Tập dữ liệu Salary Data (training set).
  + Tập dữ liệu Salary Data (training set) với outcome đã được chuẩn hóa.

Tính giá trị hàm đánh giá $R^2$ của 2 mô hình trên các tập dữ liệu tương ứng.

# Hàm đánh giá R Squared

| | Linear Regression | Decision Tree |
|---|---|---|
| Salary Data (Training set) | 0.94 | 0.97 |
| Salary Data (Training set + Scale Outcome) | 0.94 | 0.97 |

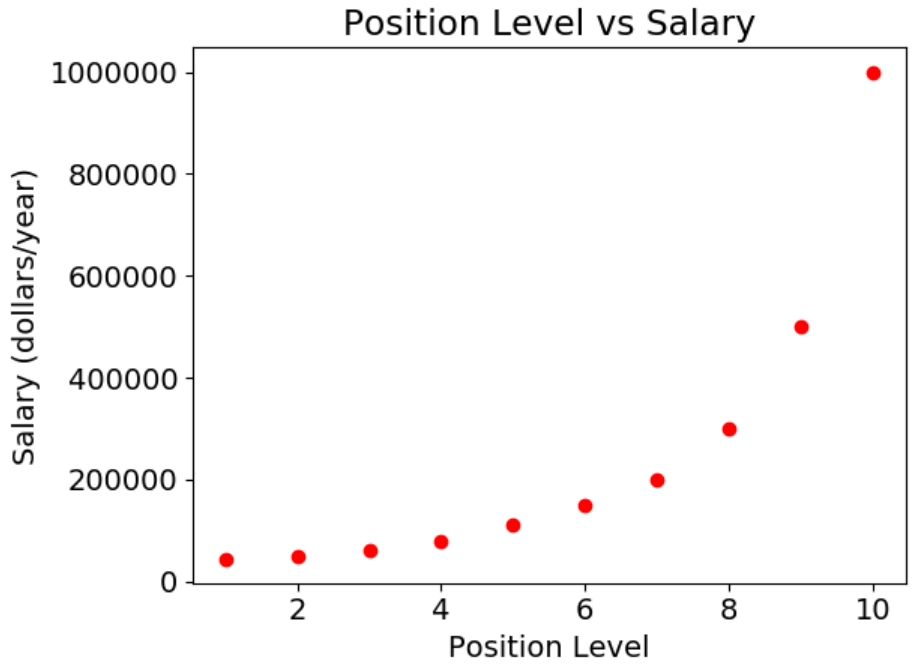
# Hàm đánh giá R Squared

- Giới thiệu lại tập dữ liệu Position Salaries.
- Mỗi điểm dữ liệu mô tả mức lương khi biết cấp độ/vị trí công việc của một người.

| Position | Level | Salary |
|---|---|---|
| Business Analyst | 1 | 45,000 |
| Junior Consultant | 2 | 50,000 |
| Senior Consultant | 3 | 60,000 |
| Manager | 4 | 80,000 |
| Country Manager | 5 | 110,000 |

# Hàm đánh giá R Squared

- Tập dữ liệu Position Salaries.
- Mỗi điểm dữ liệu mô tả mức lương khi biết cấp độ/vị trí công việc của một người.



# Hàm đánh giá R Squared

– Giới thiệu tập dữ liệu Position Salaries kèm theo một đặc trưng nhiễu.

| Noise | Level | Salary |
|---|---|---|
| 1 | 1 | 45,000 |
| 2 | 2 | 50,000 |
| 2 | 3 | 60,000 |
| 1 | 4 | 80,000 |
| 2 | 5 | 110,000 |

# Hàm đánh giá R Squared

– Ví dụ 02: Huấn luyện mô hình Linear Regression và Polynomial Linear Regression trên:

+ Tập dữ liệu Position Salaries.

+ Tập dữ liệu Position Salaries được thêm một đặc trưng nhiễu.

Tính giá trị hàm đánh giá $R^2$ của 2 mô hình trên các tập dữ liệu tương ứng.

# Hàm đánh giá R Squared

| | Linear Regression | Polynomial Regression |
|---|---|---|
| Position Salaries | 0.67 | 0.91621 |
| Position Salaries (with noise) | 0.68 | 0.91624 |

# HÀM ĐÁNH GIÁ ADJUSTED R SQUARED

# Hàm đánh giá Adjusted R Squared

– Hàm đánh giá *Adjusted R Squared*, ký hiệu là $R^2_{adjusted}$, được định nghĩa như sau:

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

+ Trong đó:

- $R^2$ là giá trị của hàm đánh giá *R Squared* trên tập dữ liệu đang xét.
- $n$ là số lượng điểm dữ liệu trong tập dữ liệu đang xét.
- $p$ là số lượng đặc trưng đầu vào.

# Hàm đánh giá Adjusted R Squared

- Hàm đánh giá *Adjusted R Squared* đã giải quyết được hai nhược điểm của hàm *R Squared*:
  - + Nếu một đặc trưng được thêm vào không cải tiến mô hình ở mức đáng kể, giá trị hàm đánh giá *Adjusted R Squared* sẽ giảm.
  - + Số lượng điểm dữ liệu trong tập dữ liệu càng lớn, giá trị của hàm *Adjusted R Squared* sẽ càng ổn định hơn, dần không phụ thuộc vào số lượng điểm dữ liệu nữa.

# Hàm đánh giá Adjusted R Squared

- Ví dụ: Huấn luyện mô hình Linear Regression và Polynomial Linear Regression trên:
  + Tập dữ liệu Position Salaries.
  + Tập dữ liệu Position Salaries được thêm một đặc trưng nhiễu.

Tính giá trị hàm đánh giá $R^2_{adjusted}$ của 2 mô hình trên các tập dữ liệu tương ứng.

# Hàm đánh giá Adjusted R Squared

| | Linear Regression | Polynomial Regression |
|---|---|---|
| Position Salaries | 0.63 | 0.87 |
| Position Slaries (with noise) | 0.62 | 0.85 |

# Chúc các bạn học tốt
# Thân ái chào tạm biệt các bạn

# ĐẠI HỌC QUỐC GIA TP.HCM
# TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN TP.HCM
# TOÀN DIỆN – SÁNG TẠO – PHỤNG SỰ