POLYNOMIAL REGRESSION

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

DATASET

Dataset

- Tên dữ liệu: Position Salaries.
- Nguồn: https://www.superdatascience.com/pages/machine-learning.
- Tập dữ liệu gồm 10 điểm dữ liệu, mỗi điểm dữ liệu gồm 3 thuộc tính, gồm:
  + Vị trí công việc (Position): mô tả tên một công việc.
  + Cấp bậc (Level): là một số nguyên trong khoảng 1 – 10, tương ứng với vị trí cao hay thấp trong một công ty.
  + Mức lương (Salary): là một số thực dương.

Dataset

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

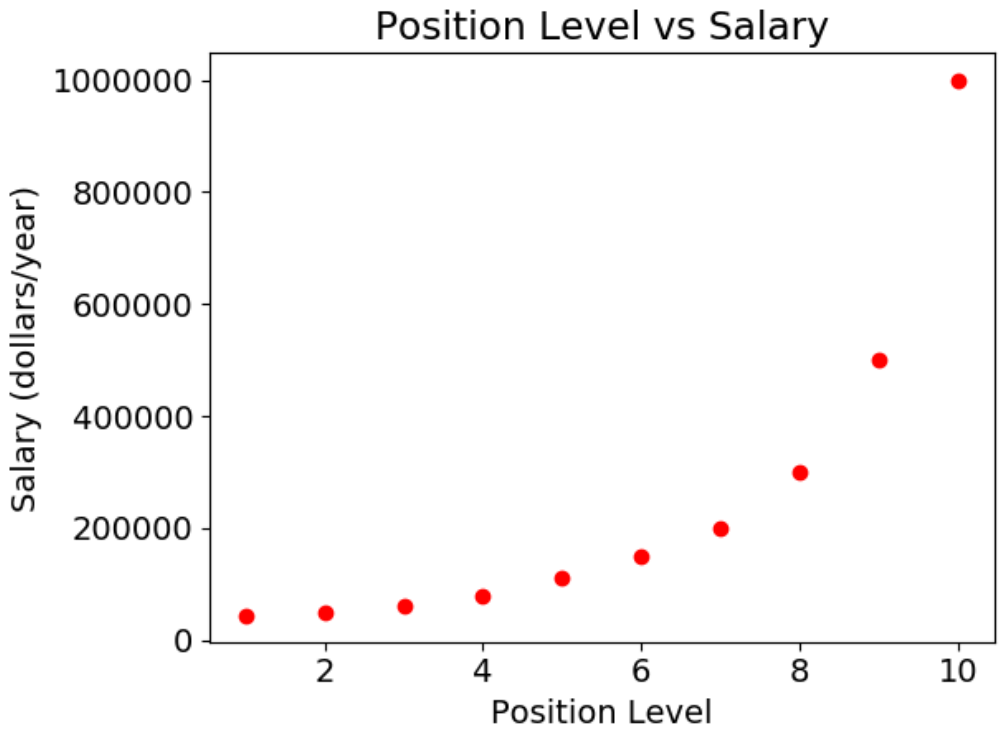
Position	Level	Salary
Region Manager	6	150,000
Partner	7	200,000
Senior Partner	8	300,000
C-level	9	500,000
CEO	10	1,000,000

Dataset

- Bài toán: Dự đoán mức lương của một người khi biết được cấp độ (vị trí) công việc của người đó.
- Ta sẽ sử dụng đồng thời thuật toán Linear Regression và thuật toán Polynomial Linear Regression cho tập dữ liệu này để so sánh hiệu suất của cả hai mô hình.

TRỰC QUAN HÓA DỮ LIỆU

Trực quan hóa dữ liệu



Trực quan hóa dữ liệu

– Đọc dữ liệu từ file csv và phân tách các giá trị đầu vào – ký hiệu là X, và giá trị đầu ra – ký hiệu là Y.

```
1. import pandas as pd
2. import numpy as np
3. dataset = pd.read_csv("Position_Salaries.csv")

4. X = dataset.iloc[:, 1:-1].values
5. Y = dataset.iloc[:, -1].values
```

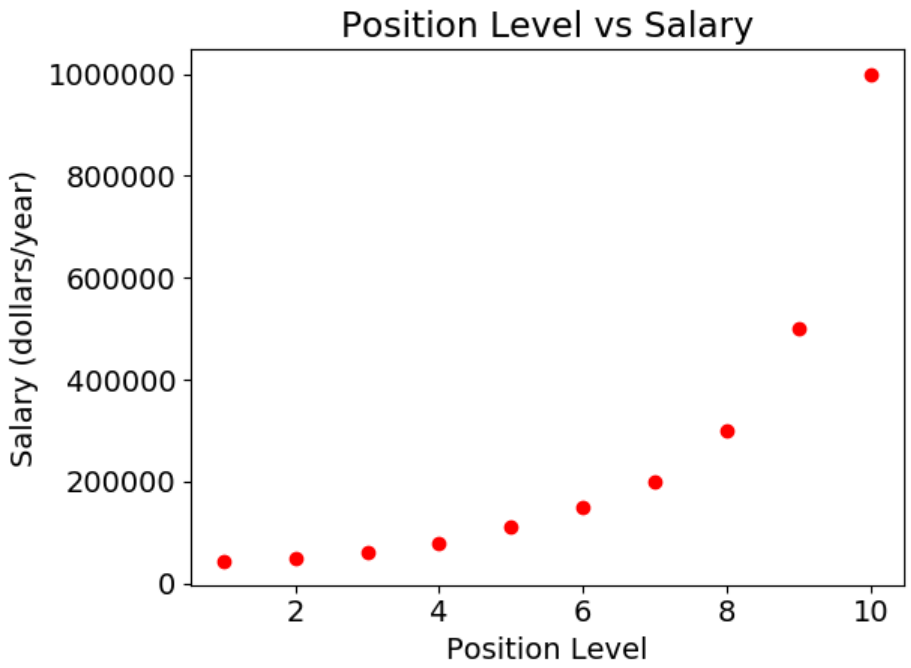
Trực quan hóa dữ liệu

– Ta vẽ các điểm (level, salary) lên mặt phẳng tọa độ để xem xét sự tương quan giữa cấp độ công việc và mức lương.

```
6. import matplotlib.pyplot as plt

7. plt.scatter(X, Y, color = "red")
8. plt.title("Position Level vs Salary")
9. plt.xlabel("Position Level")
10.plt.ylabel("Salary (dollars/year)")
11.plt.show()
```

Trực quan hóa dữ liệu



- Tập dữ liệu này không tuyến tính (không có dạng một đường thẳng).
- Do đó, thuật toán hồi quy tuyến tính – Linear Regression sẽ không hoạt động tốt trên tập dữ liệu này.

POLYNOMIAL LINEAR REGRESSION

Polynomial Linear Regression

- Polynomial Regression (hay Polynomial Linear Regression) là mô hình hồi quy đa thức.

- Mô hình Polynomial Regression đơn biến có dạng như sau:

$$y = w_0 + w_1 \times x + w_2 \times x^2 + \cdots + w_n \times x^n$$

- Trong đó:
  + y là kết quả đầu ra (outcome) hay biến phụ thuộc.
  + x là đặc trưng đầu vào (input feature) hay biến độc lập.
  + $w_0, w_1, \ldots, w_n$ là các tham số (parameters) mô hình.
  + n được gọi là bậc (degree) của mô hình.

Polynomial Linear Regression

- Mặc dù về mặc trực quan, mô hình này biểu diễn một đường cong (phi tuyến), nhưng nó vẫn được coi là một mô hình hồi quy tuyến tính đa biến.
- Từ "tuyến tính" ám chỉ mối quan hệ giữa các trọng số w với y, không phải mối quan hệ x với y.
- Mô hình Polynomial Regression đơn biến có dạng như sau:

$$y = w_0 + w_1 \times x + w_2 \times x^2 + \cdots + w_n \times x^n$$

TIỀN XỬ LÝ DỮ LIỆU

Tiền xử lý dữ liệu

- Để huấn luyện mô hình Polynomial Linear Regression, ta sẽ tính trước các biến $x, x^2, x^3, \ldots, x^n$, sau đó đưa các biến này vào huấn luyện ở mô hình Linear Regression.
- Ta dùng lớp `PolynomialFeatures` ở module `preprocessing`, package `sklearn` cho phép biến đổi trên.
- n (degree) ở bài này được đặt là 4, tức ta sẽ tính x, x^2, x^3, x^4.

```
12.from sklearn.preprocessing import PolynomialFeatures
13.poly_transform = PolynomialFeatures(degree=4)
14.X_poly = poly_transform.fit_transform(X)
```

HUẤN LUYỆN MÔ HÌNH

Huấn luyện mô hình

– Trước tiên, ta huấn luyện tập dữ liệu với mô hình Linear Regression bằng cách sử dụng lớp `LinearRegression` trong module `sklearn.linear_model`.

```
15.from sklearn.linear_model import LinearRegression
16.lin_reg = LinearRegression()
17.lin_reg.fit(X, Y)
```

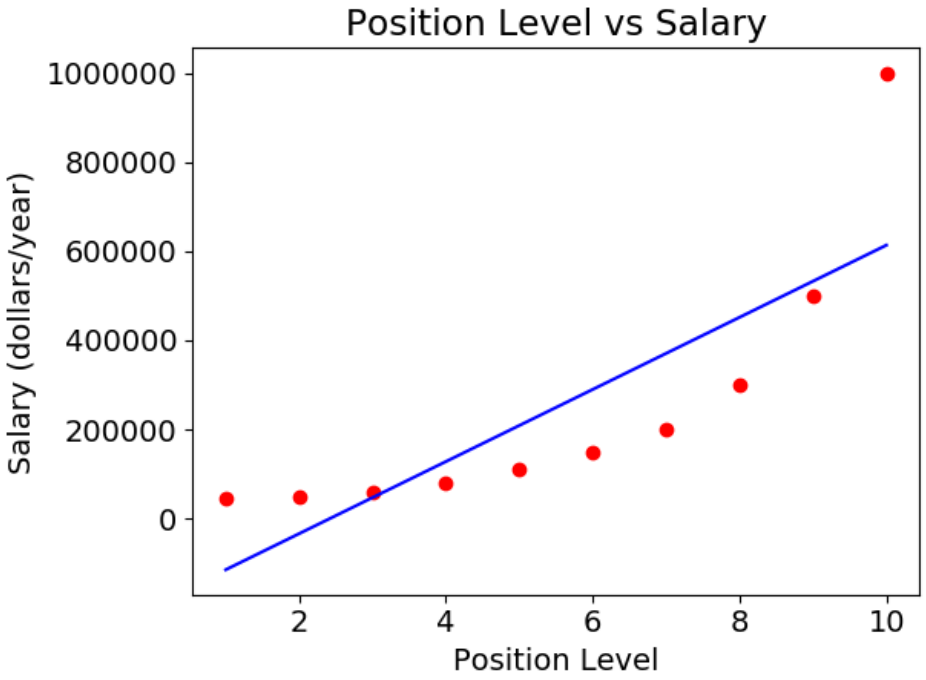
Huấn luyện mô hình

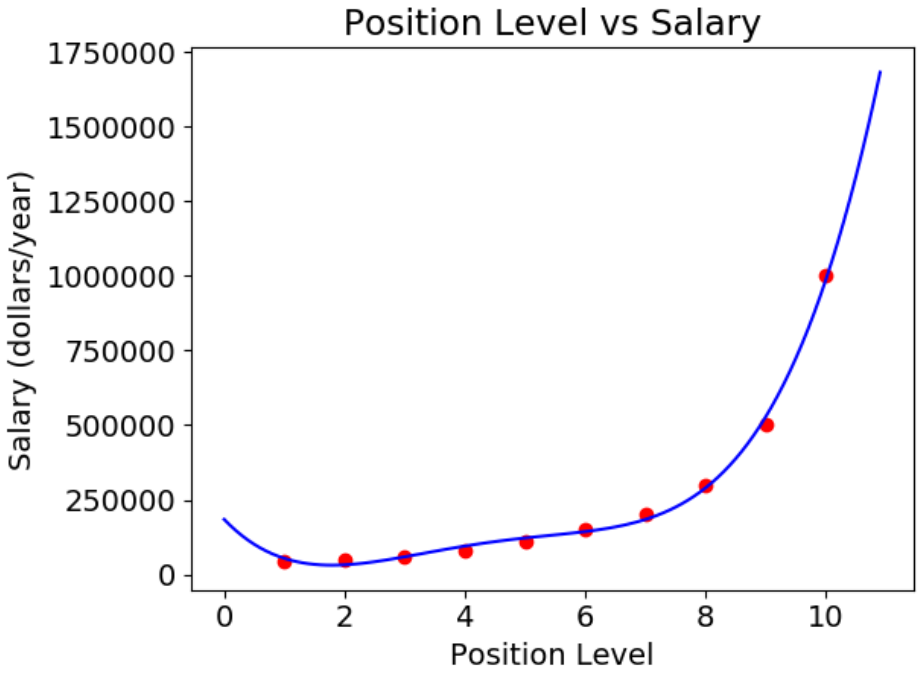
- Ta tiếp tục huấn luyện tập dữ liệu với Polynomial Linear Regression bằng cách đưa dữ liệu đã biến đổi bằng phép Polynomial Transform vào huấn luyện ở mô hình Linear Regression.

```
18.poly_lin_reg = LinearRegression()
19.poly_lin_reg.fit(X_poly, Y)
```

TRỰC QUAN HÓA KẾT QUẢ

Trực quan hóa kết quả



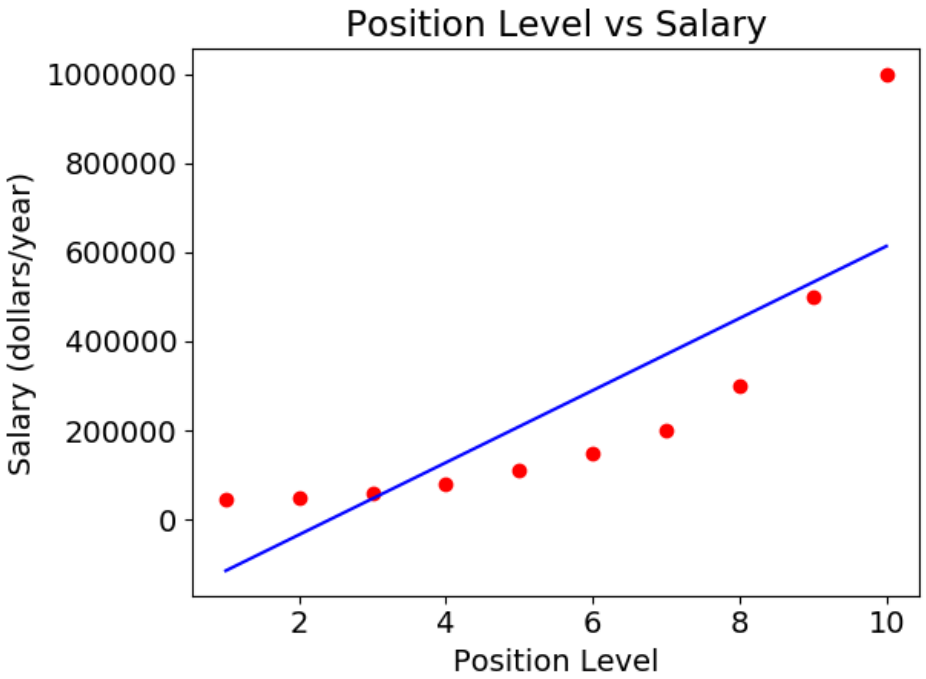


Trực quan hóa kết quả

– Trực quan hóa kết quả khi sử dụng mô hình Linear Regression.

```
24.Y_pred = lin_reg.predict(X)
25.plt.scatter(X, Y, color = "red")
26.plt.plot(X, Y_pred, color = "blue")
27.plt.title("Position Level vs Salary")
28.plt.xlabel("Position Level")
29.plt.ylabel("Salary (dollars/year)")
30.plt.show()
```

Trực quan hóa kết quả



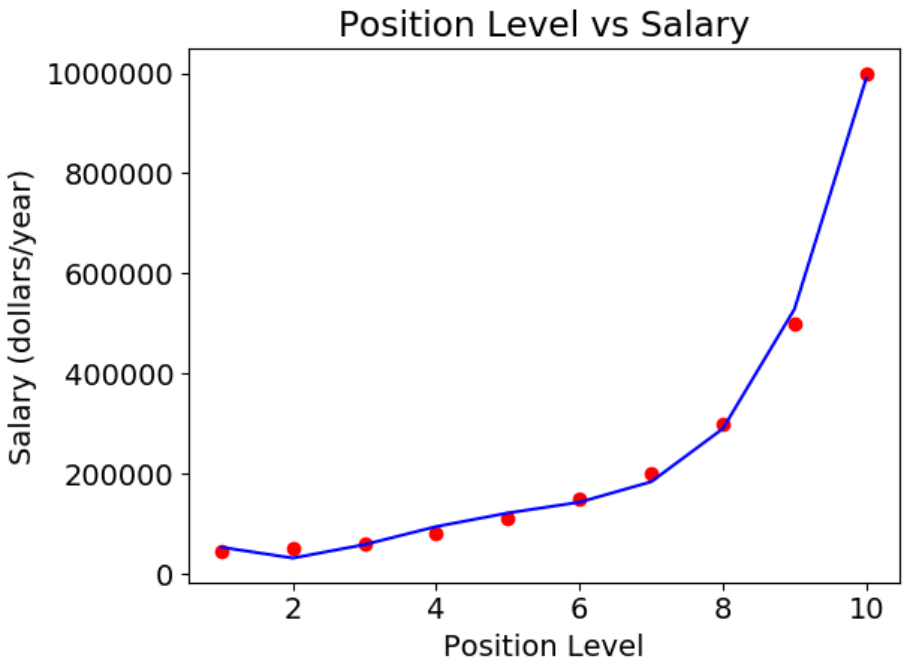
- Như đã dự đoán, mô hình Linear Regression hoạt động không tốt trên tập dữ liệu này.

Trực quan hóa kết quả

– Trực quan hóa kết quả khi sử dụng mô hình Polynomial Linear Regression.

```
31.Y_poly_pred = poly_lin_reg.predict(X_poly)
32.plt.scatter(X, Y, color = "red")
33.plt.plot(X, Y_poly_pred, color = "blue")
34.plt.title("Position Level vs Salary")
35.plt.xlabel("Position Level")
36.plt.ylabel("Salary (dollars/year)")
37.plt.show()
```

Trực quan hóa kết quả



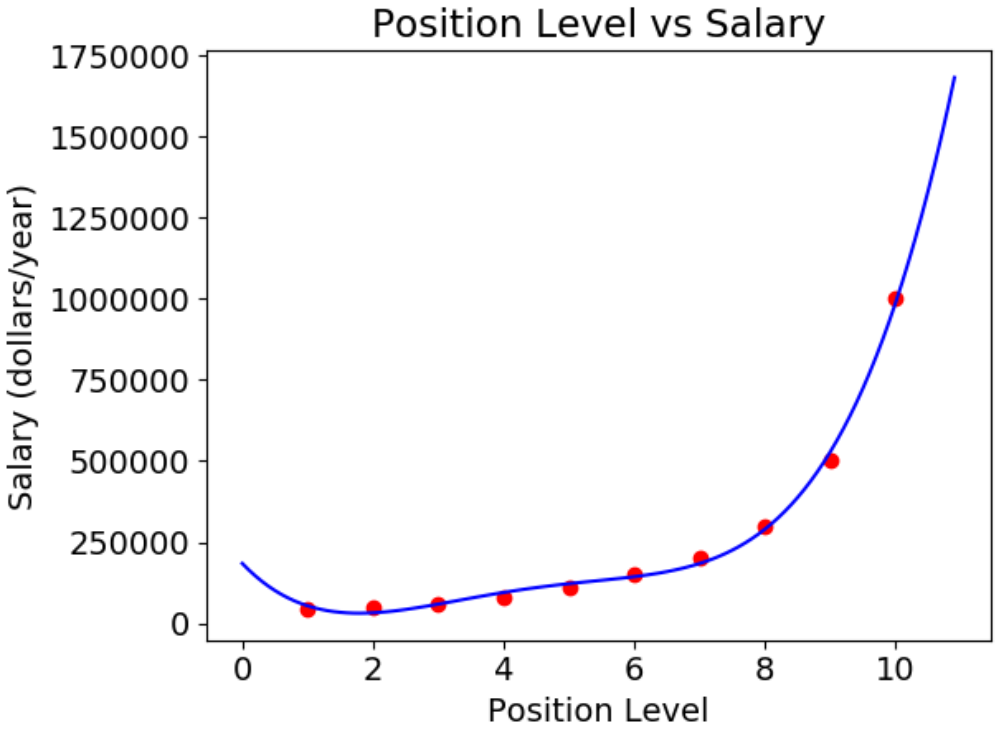
- Ta thấy mô hình Polynomial Linear Regression phù hợp với tập dữ liệu này hơn Linear Regression thông thường.

Trực quan hóa kết quả

– Vẽ lại đồ thị mà mô hình Polynomial Linear Regression dự đoán.

```
38.X_dummy = np.arange(0, 11, 0.1).reshape(-1, 1)
39.X_dummy_poly = poly_transform.transform(X_dummy)
40.Y_dummy_poly_pred = poly_lin_reg.predict(X_dummy_poly)
41.plt.scatter(X, Y, color = "red")
42.plt.plot(X_dummy, Y_dummy_poly_pred, color = "blue")
43.plt.title("Position Level vs Salary")
44.plt.xlabel("Position Level")
45.plt.ylabel("Salary (dollars/year)")
46.plt.show()
```

Trực quan hóa kết quả



Trực quan hóa kết quả

– Xây dựng hàm so sánh kết quả trên từng điểm dữ liệu.

```
47.def compare(i_example):
48.     x = X[i_example : i_example + 1]
49.     x_poly = poly_transform.transform(x)
50.     y = Y[i_example]
51.     y_pred = poly_lin_reg.predict(x_poly)
52.     print(x, y, y_pred)
```

Trực quan hóa kết quả

- Gọi hàm so sánh kết quả trên tất cả điểm dữ liệu trong tập training.

```
53.for i in range(len(X)):
54.    compare(i)
```

Trực quan hóa kết quả

Position	Level	Salary	Predicted Salary
Business Analyst	1	45,000	53,356
Junior Consultant	2	50,000	31,759
Senior Consultant	3	60,000	94,632
Manager	4	80,000	121,724
Country Manager	5	110,000	143,275

Trực quan hóa kết quả

Position	Level	Salary	Predicted Salary
Region Manager	6	150,000	184,003
Partner	7	200,000	184,003
Senior Partner	8	300,000	289,994
C-level	9	500,000	528,694
CEO	10	1,000,000	988,916

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