ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO ĐỒ ÁN 01

Bài toán Khí hậu

Môn học: Toán ứng dụng và thống kê

Sinh viên thực hiện

Bùi Quốc Trung (20120023)
Dũ Quốc Huy (20120101)
Nguyễn Kông Đại (20120448)

Giảng viên hướng dẫn

Nguyễn Đình Thức
Nguyễn Văn Quang Huy

Tháng 4 năm 2022

Mục Lục

1. Đặt vấn đề

Bài toán: Ở phần này, nhóm chúng em xin chọn nhiệt độ làm đối tượng cho bài toán. Từ đối tượng được lựa chọn, đặt ra câu hỏi "Nhiệt độ trung bình của ngày tiếp theo sẽ như thế nào?". Lấy dạng toán dự đoán làm hệ quy chiếu, tiến hành dự đoán nhiệt độ trung bình của ngày tiếp theo. Từ đó, xác định mối liên hệ giữa biến nhiệt độ trung bình của 2 ngày kế tiếp nhau. Để giải quyết vấn đề, nhóm sử dụng các biến liên quan đến nhiệt độ và chọn lấy dữ liệu trong phạm vi thành phố Hồ Chí Minh trong khoảng thời gian bắt đầu từ ngày 01/01/2002 đến ngày 30/03/2022.

2. Thu thập và xử lý dữ liệu

2.1. Thu thập

Từ đối tượng và phạm vi đã được nêu trong phần Đặt vấn đề, nhóm cần tìm dữ liệu về nhiệt độ trung bình hằng ngày tại thành phố Hồ Chí Minh trong khoảng thời gian nói trên dựa vào nguồn dữ liệu của Cơ quan Quản lí Khí quyển và Đại dương Quốc gia (NOAA) của Mỹ: https://www.ncdc.noaa.gov/cdo-web/datasets. Quá trình thu thập dữ liệu của nhóm được thực hiện theo từng bước sau:

- Bước 1: Từ trang dữ liệu của NOAA, chọn 'Tóm tắt hàng ngày' và chọn tiếp 'Công cụ tìm kiếm'.
- Bước 2: Chọn dữ liệu theo nhu cầu: thời gian từ 01/01/2002 đến 30/03/2022, tìm kiếm theo thành phố với cụm từ để tìm là 'Ho Chi Minh City'.
- Bước 3: Một bản đồ sẽ hiện ra kèm theo các kết quả tìm kiếm theo cụm từ đứng đầu là Ho Chi Minh City, VM. Chọn thêm vào giỏ hàng.
- Bước 4: Vào giỏ hàng để tiếp tục quá trình lựa chọn dữ liệu. Vì nhóm muốn chương trình đọc dữ liệu từ file csv nên sẽ lựa chọn định dạng đầu ra có đuôi csv. Lựa chọn lại thời gian như đã nếu ở Bước 2.
- Bước 5: Lựa chọn kiểu dữ liệu, ở đây nhóm chỉ cần dữ liệu về nhiệt độ trung bình nên chỉ chọn 'Nhiệt độ trung bình(TAVG)'.
- Bước 6: Xem xét lại dữ liệu theo yêu cầu, nhập địa chỉ email để nhận dữ liệu và tích chọn gửi. Trong một ít phút sẽ có một mail gửi về báo dữ liệu đã hoàn tất để có thể tải về sử dụng.

2.2. Xử lý

Bộ dữ liệu tải về gồm có 5 cột là STATION (mã trạm khí tượng), Name (Tên trạm khí tượng), DATE (ngày) và TAVG (Nhiệt độ trung bình theo °F). Theo đối tượng của bài toán chúng ta cần quan tâm đến nhiệt độ trung bình và thường ở nước ta thưởng sử dụng theo °C nên chúng ta cần chuyển cột nhiệt độ (TAVG) từ °F về °C theo công thức sau $\frac{5}{9} * (°F - 32)$

3. Phân tích, đánh giá và kết luận

3.1. Phân tích

Vấn đề của bài toán là cần quan tâm đến mối liên hệ giữa biến nhiệt độ của hai ngày kế tiếp nhau. Và với dạng bài toán dự đoán này nhóm chọn mô hình hồi quy tuyến tính để giải quyết. Bởi dựa vào đường thẳng hồi quy và phương trình đường thẳng chúng ta có thể dự đoán được biến y (nhiệt độ trung bình) khi đã có x (nhiệt độ trung bình trong ngày trước ngày của y).

Cơ bản về Hồi quy tuyến tính:

Hồi quy tuyến tính (Linear Regression) được phát triển thành mô hình hồi quy tuyến tính – LRM (Linear Regression Model) là 1 trong công cụ quan trọng trong Kinh tế lượng và là phương pháp thống kê giúp hồi quy và dự báo dữ liệu theo thuật toán giữa một giá trị liên tục với một hoặc nhiều các giá trị liên tục, định danh hay phân loại có liên quan. Hiểu 1 cách đơn giản thì Hồi quy tuyến tính là phương pháp tiếp cận tuyến tính để dự đoán biến phụ thuộc y (biến kết cục) trên trục tung Y dựa trên các biến độc lập X (biến giải thích) trên trục hoành X trong mô hình.

Phương trình hồi quy tuyến tính mà nhóm sử dụng:

$$y = B0 + B1 * X$$

Trong đó,

- y: Biến phụ thuộc
- X: Biến độc lập
- B0: Hệ số biểu diễn điểm cắt của đường thẳng hồi quy với trục Y
- B1: Hệ số góc biểu diễn độ dốc của đường hồi quy

Từ phương trình trên ta vẽ được **đường hồi quy tuyến tính** có thể tạo được sự phân bố gần nhất với hầu hết các điểm dữ liệu. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.

Qua lý thuyết cơ bản về hồi quy tuyến tính đã nêu ở trên ta thấy mô hình này phù hợp với vấn đề bài toán đã đặt ra. Vì vậy, chúng ta sẽ áp dụng mô hình hồi quy tuyến tính và sử dụng thư viện máy học phổ biến nhất cho Python là **Scikit-Learn (Sklearn)** để giải quyết bài toán này.

3.2. Đánh giá

Chúng ta lấy ngẫu nhiên 10% từ bộ dữ liệu (tức là 739 dòng dữ liệu) làm tập test để chạy thử nghiệm mô hình. Ta train các dữ liệu còn lại và tìm hiệu được hệ số hồi quy như sau:

```
B0 = [5.93783239]
B1 = [[0.78582663]]
```

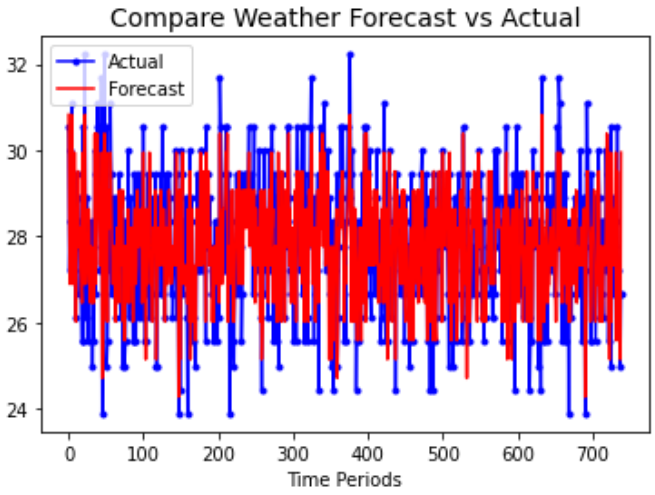
Sau đó ta chạy chạy thử nghiệm mô hình với bộ test đã lấy ở trên và thu được bảng kết quả:

	Actual	Predicted
0	30.555556	30.822342
1	27.222222	28.202920
2	28.333333	26.893209
3	30.000000	28.639491
4	31.111111	30.822342
...	...	...
734	26.666667	26.893209
735	27.222222	26.456639
736	26.666667	27.329780
737	25.000000	25.146928
738	26.666667	29.949202

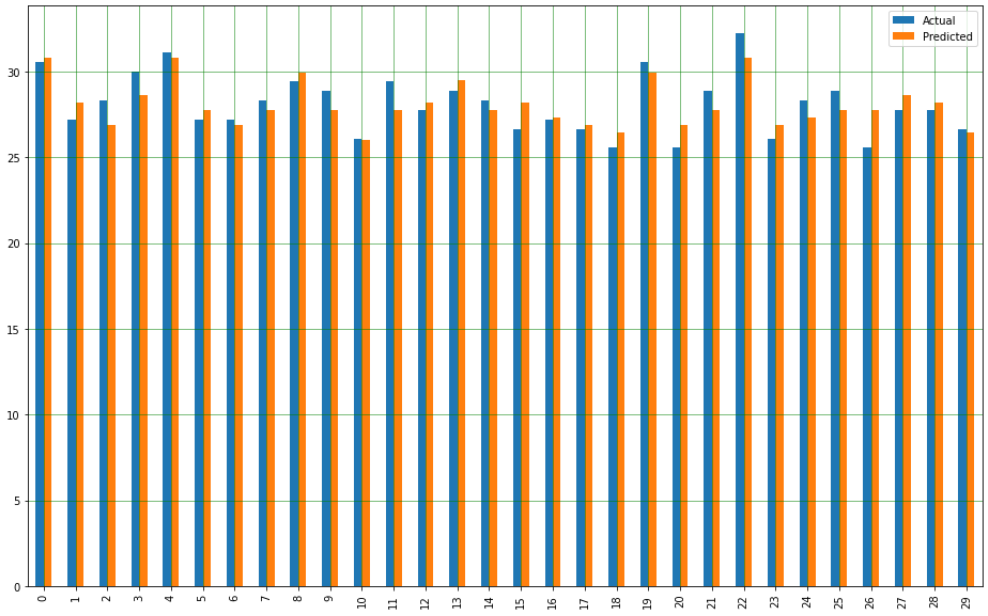
739 rows × 2 columns

Hình 1: Bảng kết quả dự đoán

Nhìn vào bảng kết quả trên ta thấy dự đoán cũng khá là chính xác. Kết quả dự đoán không lệch quá nhiều so với thực tế.Để tường minh ta có 2 biểu đồ về kết quả dự đoán này:

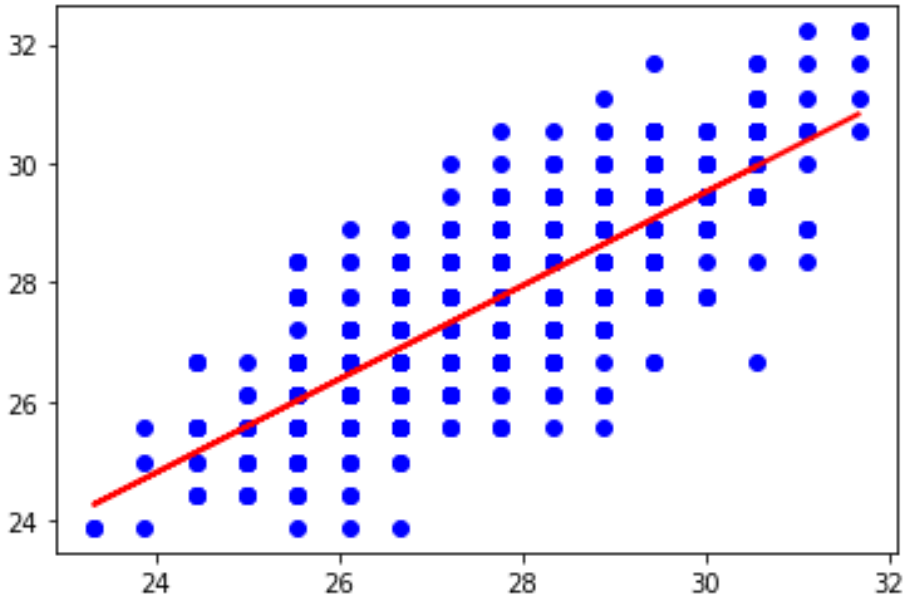


Hình 2: Biểu đồ đường về kết quả dự đoán



Hình 3: Biểu đồ cột về kết quả dự đoán cho 30 ngày

Đường hồi quy tuyến tính:



Hình 4: Đường hồi quy tuyến tính

Từ hình trên ta thấy các điểm phân bố cũng là gần đường hồi quy nên khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó sẽ không quá cao điều này làm cho độ chính xác của dự đoán chúng ta tương đối chính xác.

Các số liệu để đánh giá độ hiệu quả mô hình hồi quy đã áp dụng trong bài toán:

```
R Squared: 0.6285466543401419
Mean Squared Error: 0.9225370329161722
Root Mean Squared Error: 0.9604879139875588
Mean Absolute Error: 0.753348221452829
```

R Squared (R^2): cho biết mô hình của chúng ta hợp với dữ liệu ở mức 62.85%. Và ta có thể thấy các giá trị sai số (MSE), (RMSE) và (MAE) đều nhỏ hơn 1, điều này cho thấy thuật toán của nhóm chọn là hợp lý và đưa ra các dự đoán khá chính xác.

3.3. Kết luận

Thuật toán mà nhóm đã chọn để giải quyết vấn đề đã đặt ra là chính xác. So với các mô hình nhóm đã từng thử nghiệm thừ mô hình hồi quy tuyến tính này là tốt và dự đoán hợp lý nhất. Thời gian chạy của chương trình cũng khá nhanh và mô hình này khá là dễ hiểu. Tuy nhiên vẫn còn một số dự đoán còn hơi lệch nhưng chiếm không nhiều. Nhìn chung ta có thể thấy mô hình hồi quy tuyến tính này khá phù hợp cho các dạng bài toán dự đoán.

Tài liệu tham khảo

[1] Dữ liệu về nhiệt độ: https://www.ncdc.noaa.gov/cdo-web/datasets

[2] Hồi quy tuyến tính trong Python với Scikit-Learn: https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html?fbclid=IwAR3Zz-fqjiJ2vxs0zjH5Qp2uWLnlbicjPvEMH6L9rz06liEKnpbWpyDSncE

[3] Linear Regression - Hồi quy tuyến tính trong Machine Learning: https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3?fbclid=IwAR2TCyGCefwIMjM55iQ3ZULANcWginKov2ie1jU6E7VdjLSxGMXxQf1d60k