Dự đoán nhiệt độ điểm sương trong không khí bằng mô hình hồi qui tuyến tính

Trịnh Quốc Huy - 20120013
Nguyễn Anh Khoa-20120118
Võ Thị Phước Thảo - 20120191

Ngày 9 tháng 4 năm 2022

Giảng viên hướng dẫn: Nguyễn Đình Thúc - Nguyễn Văn Quang Huy

1 Giới thiệu

Theo định nghĩa về khí tượng thủy văn, nhiệt độ điểm sương (Dew point) là nhiệt độ mà tại đó hình thành sự ngưng tụ hơi nước. Điểm sương là thước đo lượng hơi nước có trong chất khí và thường được dùng rất nhiều trong các lĩnh vực liên quan đến thời tiết như để kiểm tra, dự đoán lượng mưa, dự đoán các hiện tượng thời tiết.
Theo định luật Dalton về áp suất không khí, trong không khí tồn tại ba thành phần chính là: Ni tơ, hơi nước và khí oxi. Vì thế mà áp suất xấp sỉ của không khí chính là tổng áp suất của ba khí nói trên và cụ thể hơn áp suất của không khí chính là tổng của ba áp suất thành phần [1].Ngoài ra, định luật Dalton còn chỉ ra rằng áp suất riêng của hơi nước là một hàm của nhiệt độ, trong điều kiện này, việc thêm nhiều hơi nước sẽ dẫn đến việc hình thành ngưng tụ. Hiện tượng ngưng tụ này có thể được khai thác để đo hàm lượng hơi nước và nhờ đi qua bề mặt kiểm soát nhiệt độ được làm lạnh khiến khối hơi nước này ngưng tụ lại, tạo thành điểm sương và nhiệt độ khi hình thành sự ngưng tụ được gọi là nhiệt độ điểm sương. Nhiệt độ điểm sương là một yếu tố quan trọng trong việc dự đoán thời tiết, đồng thời là một thang đo để dự đoán các hiện tượng tự nhiên có thể gây nguy hiểm cho con người, nhất là trong tình trạng biến đổi khí hậu và hiệu ứng nhà kính như hiện nay, đó là lí do vì sao nhóm chọn đề tài này vì có tầm ảnh hưởng rất lớn tới khí hậu Việt Nam nói riêng và thế giới nói chung [2].
Trong bài báo cáo lần này, nhóm đề xuất phương pháp sử dụng mô hình hồi qui tuyến tính (Linear Regression) để thực hiện công việc dự đoán nhiệt độ điểm sương tại một nhiệt độ và một áp suất bất kì dựa trên hai yếu tố là nhiệt độ hiện tại và áp suất qua từng ngày do đặc tính ngưng tụ cần tốn khá nhiều thời

gian nên nhóm chọn mốc thời gian theo ngày. Quá trình thực hiện nhóm có đề xuất sử dụng các phương pháp dựa trên công thức nghiệm và thuật toán tối ưu Gradient Descent lên mô hình hồi qui tuyến tính. Đồng thời nhóm có tiến hành thử nghiệm bài toán trên mô hình Neural Network để so sánh các kết quả và hiệu quả của phương pháp nhóm đề xuất là sử dụng công thức nghiệm của Linear Regression.

2 Cơ sở lí thuyết

2.1 Giới thiệu mô hình hồi qui tuyến tính

Bài toán được đặt ra là chúng ta có một tập điểm dữ liệu x và y, chúng ta tìm mối quan hệ tuyến tính giữa x và y đó. Giả sử tập điểm X và Y của chúng ta có dạng là:

$$X = (x_1, x_2, x_3, ..., x_n) \tag{1}$$

$$Y = (y_1, y_2, y_3, ..., y_n) \tag{2}$$

Theo yêu cầu của bài toán, chúng ta phải tìm một hàm số F(x) để thõa điều kiện là Y = F(X) với F(x) là một hàm tuyến tính có dạng:

$$Y = w_1x_1 + w_2x2 + ... + wnxn \tag{3}$$

. Trong phương trình (3) ta có tập hợp $W = (w_1, w2, .., wn)$ là các hệ số chúng ta cần phải tối ưu trong bài toán này [3].

2.2 Sai số dự đoán

Ta gọi giá trị dự đoán của mô hình là $\hat{y}$, và giá trị thật của dữ liệu là y, từ đó dựa trên công thức thống kê ta tính được sai số (e) của mô hình là:

$$e = |y - \hat{y}| \tag{4}$$

Tuy vậy để có thể thuận tiện cho việc giải quyết bài toán tối ưu bằng phương pháp đạo hàm, ta có sai số dự đoán của mô hình được viết lại như sau:

$$e = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - w\overline{x})^2 \tag{5}$$

Sai số này chính là để đánh giá được mức độ chính xác của mô hình khi dự đoán một điểm dữ liệu so với dữ liệu thực tế để mà từ đó có thể điều chỉnh lại các trọng số w cho phù hợp với bài toán và lời giải của bài toán [3].

2.3 Hàm mất mát

Giả sử ta có tập điểm dữ liệu với N điểm dữ liệu chạy từ 1 tới N, yêu cầu để có thể tối ưu được các trọng số w để cho tổng các sai số trong điểm dữ liệu là nhỏ nhất, từ đó chúng ta cần thiết lập hàm mất mát (Loss Function), ta gọi là L(w), cho N điểm dữ liệu dựa trên tổng sai số dự đoán với công thức sai số dự đoán (5):

$$L(w) = \frac{1}{2}\sum_{1}^{N}||y - \hat{y}||_2^2 = \frac{1}{2}||y - w\overline{x}||_2^2 = \frac{1}{2}\sum_{1}^{N}(y - \hat{y})^2 = \frac{1}{2}\sum_{1}^{N}(y - w\overline{x}_i)^2 \quad (6)$$

Để có thể tối ưu được hàm L(w), chúng ta cần tìm w sao cho L(w) đạt giá trị nhỏ nhất, điều này tương đương chúng ta đi tìm giá trị của w, ta tạm gọi là$\hat{w}$ sao cho hàm L(w) có giá trị nhỏ nhất. Ta tổng quát hóa bài toán này như sau:

$$\hat{w} = argmin_w(L(w)) \quad (7)$$

Để tìm được giá trị nhỏ nhất cho bài toán này ta sẽ tiến hành giải theo phương pháp tối ưu lồi.

2.4 Phương pháp tối ưu và công thức nghiệm

Ý tưởng của bài toán tối ưu lồi cho các mô hình Linear Regression chính là việc sử dụng các công thức đạo hàm và tìm giá trị nhỏ nhất dựa trên khoảng và các giá trị cực trị. Với hàm mất mát đã đề cập theo công thức số (6) thì để tìm được giá trị $\hat{y}$ cho hàm L(w) theo công thức số (7) thì ta sẽ tính toán cực trị tại đạo hàm bằng 0. Theo công thức số (6) ta sẽ có được đạo hàm của hàm L(w) như sau:

$$L'(w) = \frac{\partial L(w)}{\partial w} = \overline{X}^T(w\overline{X} - y) \quad (8)$$

Để tìm được giá trị w tối ưu cho phương trình trên, chúng ta cho đạo hàm của L tại w bằng 0, khi đó chúng ta thu được kết quả là:

$$\overline{X}^T\overline{X}w = \overline{X}^T y \quad (9)$$

Với phương trình (9), ta thu được kết quả w, tạm gọi là $\hat{w}$ là nghiệm tối ưu của phương trình với công thức sau khi biến đổi phương trình (9) như sau:

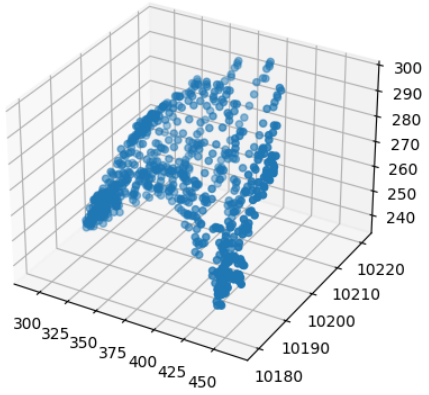
$$\hat{w} = (\overline{X}^T\overline{X})^{-1}\overline{X}^T y \quad (10)$$

Công thức (10) chính là công thức nghiệm của bài toán dựa trên mô hình hồi qui tuyến tính (Linear Regression) mà nhóm sử dụng [3].

3 Thí nghiệm

3.1 Chuẩn bị dữ liệu

Để có thể thực hiện được bài toán này, nhóm sử dụng bộ dữ liệu của NOA, là trung tâm khi thống kê về dữ liệu nhiệt độ và áp suất của trung tâm khí tượng và môi trường của Mỹ, các dữ liệu được thống kêu theo ngày bao gồm nhiệt độ và áp suất đồng thời là các điểm nhiệt độ điểm sương [1]. Bộ dữ liệu bao gồm 750 mẫu về nhiệt độ [1], áp suất và điểm sương qua từng thời điểm theo ngày tại trạm GHCND:USW00003812 của Mỹ, từ đó dựa trên định luật Dalton chúng ta có thể hiểu được dữ liệu và phân tích dữ liệu trong không gian vector để có thể hiểu hơn về dữ liệu. Sau khi chọn lọc và thống kê dữ liệu, nhóm tiến hành mô phỏng dữ liệu trên không gian vectors:



Hình 1: Hình ảnh mô tả phân phối dữ liệu

Sau khi tiến hành phân tích dữ liệu, nhóm nhận thấy được phân bố dữ liệu theo mô hình hồi qui tuyến tính nên nhóm quyết định chọn mô hình này để tiến hành thí nghiệm và dự đoán trên dữ liệu, nhóm làm sạch dữ liệu bằng cách bỏ đi các phần bị rỗng trong bộ dữ liệu. Nhóm chia tập dữ liệu theo tỉ lệ 8:2, phần 8 cho phần huấn luyện và phần 2 cho phần kiểm tra độ chính xác của mô hình.

3.2 Xây dựng mô hình

Nhóm xây dựng mô hình Linear Regression bằng 2 thư viện chính là Numpy và Pandas, nhóm tiến hành cài đặt lại Linear Regression theo thuật toán được nêu ra trong cơ sở lí thuyết. Ngoài ra, nhóm cũng tiến hành cài đặt thuật toán Linear Regression trên thư viện Scikit-learn được tối ưu bằng phương pháp đạo hàm

để so sánh giữa phương pháp sử dụng công thức nghiệm và phương pháp đạo hàm. Nhóm cũng tiến hành thử nghiệm trên Neural Network [4] với regression output để so sánh kết quả với phương pháp mà nhóm đề xuất.

3.3 Đánh giá mô hình

Nhóm đánh giá mô hình qua các thang đo, Mean Absolute Error (Mae), Mean Square Error (MSE) cho n điểm dữ liệu kiểm thử theo công thức như sau:
- Với Mae [5]:

$$MAE_{Loss} = \frac{1}{n}|\sum_{i=1}^{n}(y_i - \hat{y_i}| \tag{11}$$

- Với MSE [6]:

$$MSE_{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{12}$$

Trong đó y_i là giá trị thực tế của từng điểm dữ liệu, $\hat{y_i}$ là giá trị dự đoán của mô hình trên từng điểm dữ liệu.

4 Kết quả

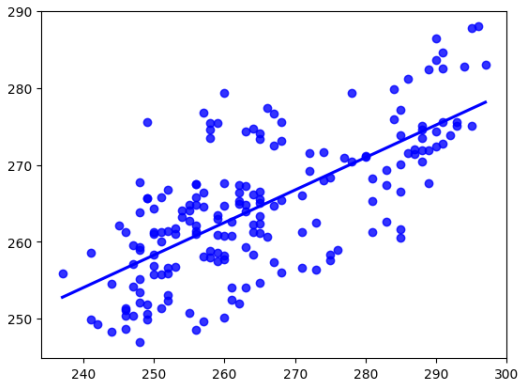
Sau khi đánh giá mô hình, nhóm thu được các kết quả như sau:

Phương pháp	MAE	MSE
Linear Regression với công thức nghiệm	7.012	42.364
Linear Regression với tối ưu bằng đạo hàm	8.921	45.526
Neural Network với tối ưu bằng đạo hàm	9.036	52.453

Bảng 1: Bảng kết quả của phương pháp nhóm đề xuất so với các phương pháp khác

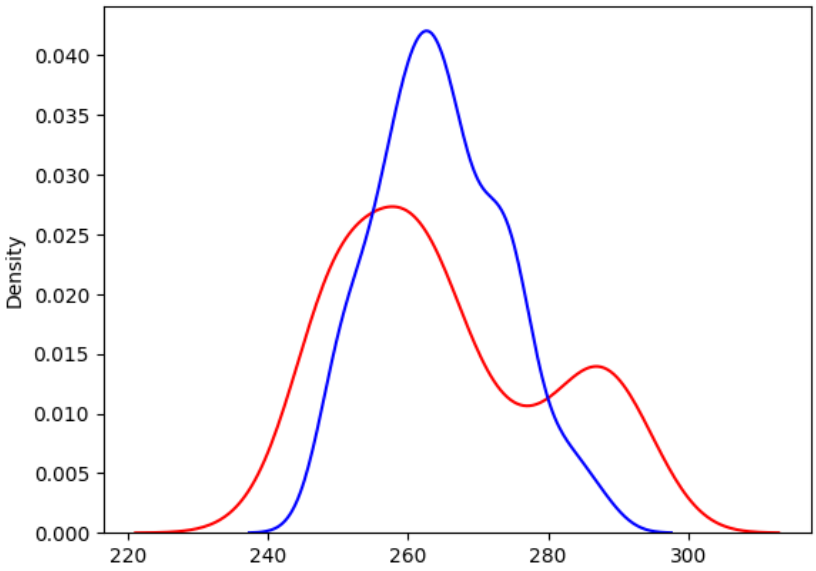
Lưu ý: Các trọng số của mô hình được nhóm cài đặt ở điều kiện tối ưu nhất và công bằng nhất trên cùng 1 tập dữ liệu.

Nhóm tiến hành phân tích mô hình mà nhóm thu được so với tập dữ liệu gốc và nhóm thu được một mô hình được biểu diễn trên đồ thị dưới hình 2D như sau:



Hình 2: Hình ảnh minh họa mô hình nhóm thu được sau quá trình huấn luyện

Nhóm so sánh các giá trị output của mô hình với giá trị thực tế để kiểm chứng độ tương quan của mô hình mà nhóm đề xuất sử dụng :



Hình 3: biểu đồ kiểm tra tương quan giữa dữ liệu dự đoán của mô hình và dữ liệu thực tế

5 Kết luận

Tóm lại, kết quả dự đoán của mô hình Linear Regression của nhóm đề xuất đạt kết quả khả quan trên tập dữ liệu 750 mẫu của bộ dữ liệu. Tuy vậy kết quả còn

một số nhược điểm khiến cho mô hình không thể đạt được kết quả tốt trên tập dữ liệu kiểm thử vì một số lí do như sau:
- Các điểm nhiễu trong bộ dữ liệu: Trong bộ dữ liệu có các điểm bị nhiễu so với phân phối chung của dữ liệu làm cho mô hình không được tốt.
- Các điểm dữ liệu còn ít: Dữ liệu trong bộ dữ liệu chưa được nhiều nên vì thế mô hình chưa thể tìm được một mô hình sát nhất trên tất cả các điểm dữ liệu.
Mô hình dựa trên công thức nghiệm đạt kết quả chính xác hơn mô hình sử dụng dựa trên phương pháp đạo hàm. Tuy vậy, nếu ứng dụng mô hình này trên một tập dữ liệu nhiều thuộc tính hơn thì sẽ gặp vấn đề về tài nguyên tính toán và bất lợi hơn phương pháp sử dụng đạo hàm để tối ưu.

Tài liệu

[1] J. C. Carman, T. Clune, F. Giraldo, M. Govett, A. Kamrath, T. Lee, D. McCarren, J. Michalakes, S. Sandgathe, T. Whitcomb, Position paper on high performance computing needs in earth system prediction, bibliography (2017).
URL https://repository.library.noaa.gov/view/noaa/14319

[2] H. Viana, P. Porto, The development of dalton's atomic theory as a case study in the history of science: Reflections for educators in chemistry, Science and Education 19 (2009) 75–90. doi:10.1007/s11191-008-9182-2.

[3] K. Kumari, S. Yadav, Linear regression analysis study, Journal of the Practice of Cardiovascular Sciences 4 (2018) 33. doi:10.4103/jpcs.jpcs_8_18.

[4] B. Mehlig, Machine Learning with Neural Networks, Cambridge University Press, 2021. doi:10.1017/9781108860604.
URL https://doi.org/10.1017%2F9781108860604

[5] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, C.-H. Lee, On mean absolute error for deep neural network based vector-to-vector regression, IEEE Signal Processing Letters 27 (2020) 1485–1489. doi:10.1109/lsp.2020.3016837.
URL https://doi.org/10.1109%2Flsp.2020.3016837

[6] C. Sammut, G. I. Webb (Eds.), Mean Squared Error, Springer US, Boston, MA, 2010, pp. 653–653. doi:10.1007/978-0-387-30164-8_528.
URL https://doi.org/10.1007/978-0-387-30164-8_528