Đồ án cuối kỳ
Nhập môn Khoa học dữ liệu

Nhóm

16

Giáo viên hướng dẫn	
Thầy: Trần Trung Kiên	**Cô**: Phan Thị Phương Uyên

DANH SÁCH THÀNH VIÊN	
Họ và tên	**MSSV**
Lê Thanh Sơn	18127199
Lê Hồng Quang	18127190

1.

Giới thiệu đề tài

- Thu thập dữ liệu về các chỉ số và chất lượng không khí tại một số nơi:
- Từ 0h ngày 1/6/2021 - 0h ngày 25/6/2021 của tỉnh Quảng Bình,Việt Nam.
- Từ 0h ngày 1/6/2021 - 0h ngày 1/9/2021 của tỉnh Quyriq, Iran.
- Từ 0h ngày 1/12/2020 - 0h ngày 20/12/2020 của TP. Vũ Hán, tỉnh Hồ Bắc, Trung Quốc.

2.

THU THẬP DỮ LIỆU

- Nguồn: https://openweathermap.org/
- Phương pháp thu thập dữ liệu : Dùng API mà trang web cung cấp
- Cú pháp:
- http://api.openweathermap.org/data/2.5/air_pollution/history?lat={lat}&lon={long}&start={start}&end={end}&appid={my_api_key}

	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi
0	313.76	0.01	0.79	36.84	0.21	2.92	3.11	0.47	1
1	320.44	0.05	0.85	40.77	0.31	2.91	3.15	0.51	1
2	333.79	0.08	0.75	49.35	0.47	3.35	3.69	0.57	1
3	357.15	0.09	0.69	62.94	0.70	4.78	5.32	0.72	1
4	397.21	0.08	0.68	82.97	1.12	7.92	8.72	0.94	2
...	...	...	...	...	...	...	...	...	...
3230	1655.58	27.49	93.22	0.00	51.02	254.16	290.67	6.40	5
3231	1735.69	32.63	90.48	0.00	61.99	259.78	299.59	8.49	5
3232	1869.20	38.89	89.11	0.00	77.25	270.65	314.25	11.15	5
3233	2376.56	58.12	87.74	0.00	103.00	314.65	371.35	15.20	5
3234	3124.24	91.20	87.74	0.07	131.61	377.03	452.87	19.25	5

3235 rows × 9 columns

3

Khám phá dữ liệu

- Dữ liệu có 3235 dòng và 9 cột
- Dữ liệu không bị lặp lại
- Ý nghĩa các cột:

- **aqi**: Chỉ số chất lượng không khí (1,2,3,4,5) với (1 = Good; 2 = Fair; 3 = Moderate; 4 = Poor ; 5 = Very Poor)
- **co**: Nồng độ CO (Carbon monoxide), µg/m³
- **no**: Nồng độ NO (Nitrogen monoxide), µg/m³
- **no2**: Nồng độ NO_2 (Nitrogen dioxide), µg/m³
- **o3**: Nồng độ O_3 (Ozone), µg/m³
- **so2**: Nồng độ SO_2 (Sulphur dioxide), µg/m³
- **pm2_5**: Nồng độ PM_{2_5} (Fine particles matter), µg/m³
- **pm10**: Nồng độ PM_{10} (Coarse particulate matter), µg/m³
- **nh3**: Nồng độ NH_3 (Ammonia), µg/m³

4.

Đặt câu hỏi và trả lời

Chất lượng không khí được phân loại như thế nào từ các chỉ số của các chất trong không khí?

Ý nghĩa thực tế của câu hỏi: Việc xác định chất lượng không khí từ các chất có trong không khí giúp phát triển thiết bị đo lường chất lượng không khí. Cảnh báo người dân các biện pháp bảo vệ bản thân và cộng đồng kịp thời như đeo khẩu trang, hạn chế sử dụng các phương tiện giao thông gây ô nhiễm... Ngoài ra, chính quyền địa phương có thể đưa ra các biện pháp nhằm hạn chế lượng khí thải từ các nhà máy, xí nghiệp, trồng thêm nhiều cây xanh để làm giảm ô nhiễm không khí, từ đó nâng cao chất lượng cuộc sống của con người.

5

Tiền xử lý

Cột output:

- Có kiểu dữ liệu dạng số
- Không có giá trị thiếu
- Tỉ lệ các lớp được phân bố khá đều nhau.

2	21.298300
4	20.340031
3	20.247295
1	19.103555
5	19.010819

TÁCH CÁC TẬP:

- Tách tập huấn luyện, validation, test theo tỉ lệ 70%: 20%: 10%

TẬP HUẤN LUYỆN:

- Tất cả các cột đều có kiểu dữ liệu số
- Các giá trị phân bổ:

	co	no	no2	o3	so2	pm2_5	pm10	nh3
missing_percentage	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
min	140.19	0.00	0.13	0.00	0.05	0.50	0.56	0.00
lower_quartile	185.20	0.00	0.50	35.40	0.20	9.10	22.10	0.40
median	217.00	0.00	0.90	68.70	0.40	14.30	39.70	0.70
upper_quartile	347.10	0.10	2.00	88.70	1.00	22.00	74.20	1.60
max	12390.14	658.04	175.48	217.44	534.06	1205.88	1395.92	19.25

TIỀN XỬ LÝ:

- Với các cột dạng số, ta sẽ điền giá trị thiếu bằng giá trị mean của cột. Với *tất cả* các cột dạng số trong tập huấn luyện, ta đều cần tính mean, vì ta không biết được cột nào sẽ bị thiếu giá trị khi dự đoán với các véc-tơ input mới.
- Cuối cùng, khi tất cả các cột đã được điền giá trị thiếu và đã có dạng số, ta sẽ tiến hành chuẩn hóa bằng cách trừ đi mean và chia cho độ lệch chuẩn của cột để giúp cho các thuật toán cực tiểu hóa như Gradient Descent, LBFGS, ... hội tụ nhanh hơn.

6

MÔ HÌNH HÓA

Mô hình MLP- Classifier

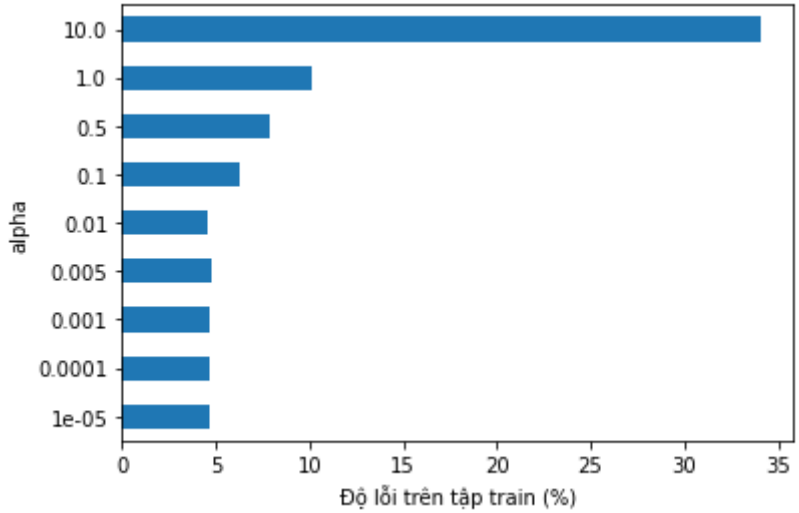
• Ta sẽ sử dụng mô hình MLP để phân lớp (với các siêu tham số hidden_layer_sizes=(20), activation='relu', solver='adam', random_state=0, max_iter=600)

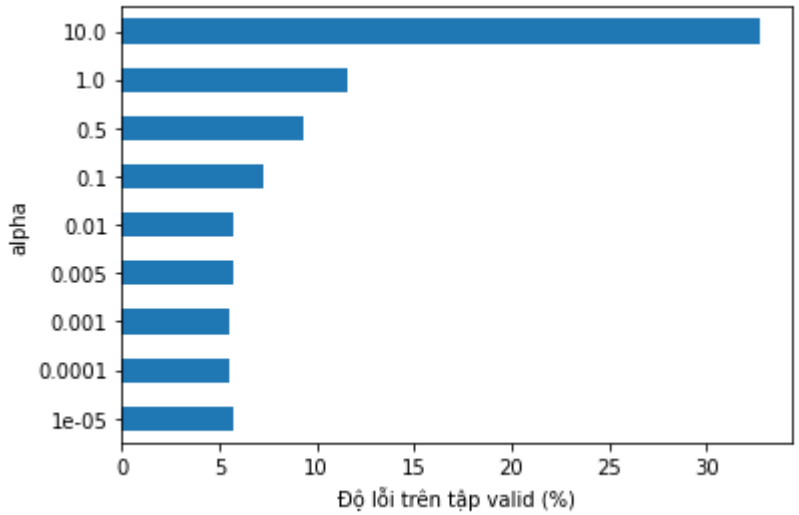
• Chọn activation=relu vì hàm relu phổ biến hiện nay và sẽ tính toán nhanh hơn hàm tanh và logistic.

• Alpha: `[0.00001, 0.0001, 0.001, 0.005, 0.01, 0.1, 0.5, 1, 10]`

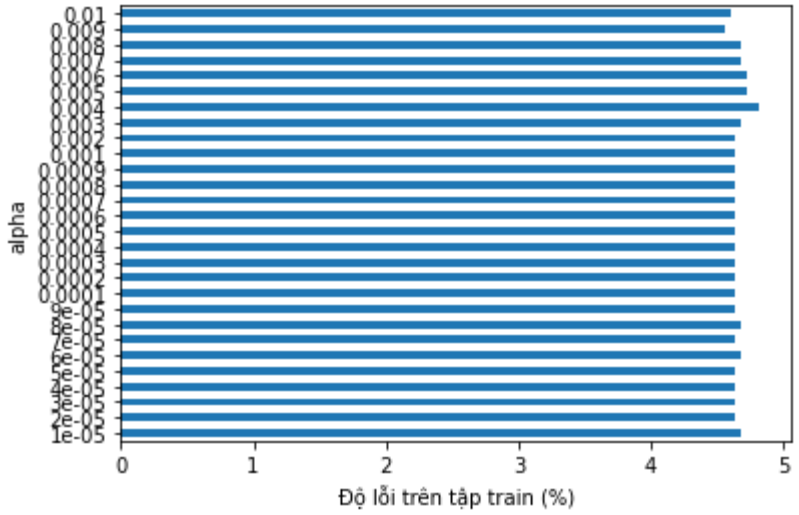
• Chọn solver = adam vì dữ liệu trên các tập dữ liệu tương đối lớn)

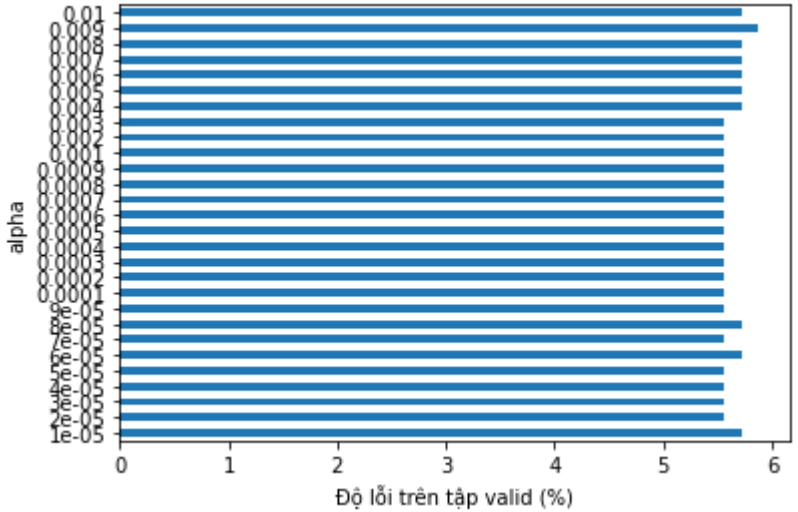
Độ lỗi trên các tập dữ liệu





Độ lỗi trên tập valid khi alpha đi từ 0.00001 - 0.01 tăng rồi lại giảm, nên tiếp tục chia nhỏ miền này để tìm alpha tốt nhất





Độ lỗi nhỏ nhất trên tập valid là 5.56%, alpha tốt nhất là 0.00002

Huấn luyện lại mô hình với cả tập train và tập valid

Với alpha = 0.00002:

- Độ lỗi trên tập train + valid: 3.81%
- Độ lỗi trên tập test: 3.4%

Nhận xét

•Mô hình MLP-Classifier fit khá tốt trên tập dữ liệu

•Độ lỗi trên cả tập (train+val) và tập test tương đối thấp

•Khi chia nhỏ alpha hơn nữa thì độ lỗi trên cả 2 tập train và valid dường như không thay đổi nhiều.

•Dự đoán trên tập train tốt hơn tập valid. Nhưng khi train model với alpha tốt nhất và dự đoán thì có sự bất ngờ, độ lỗi trên tập test tốt hơn cả độ lỗi trên tập train + val.

•Dựa vào hai biểu đồ trên , ta thấy siêu tham số alpha không tỉ lệ thuận với độ lỗi trên cả hai tập. Tại một số điểm trên miền giá trị alpha đã xét, độ lỗi trên hai tập train và valid khi đang giảm sẽ đột ngột tăng lên sau đấy giảm tiếp.

Mô hình Softmax Regression

- Độ lỗi trên tập train + valid: 14.5%
- Độ lỗi trên tập test: 13.89%

Nhận xét

•Độ lỗi trên cả hai tập khá cao (trên 10%)

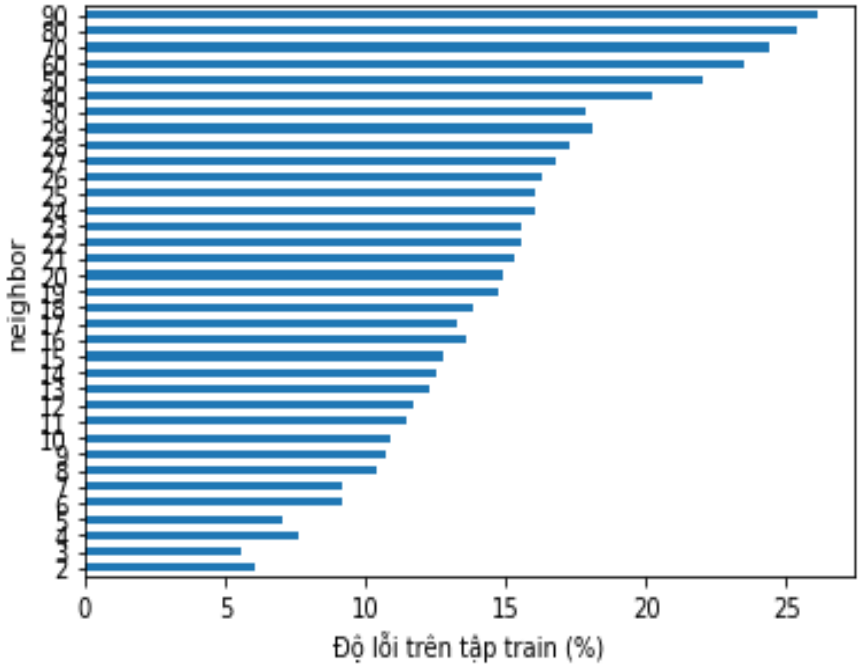
•Độ lỗi trên cả hai tập không chênh lệch nhau quá nhiều.

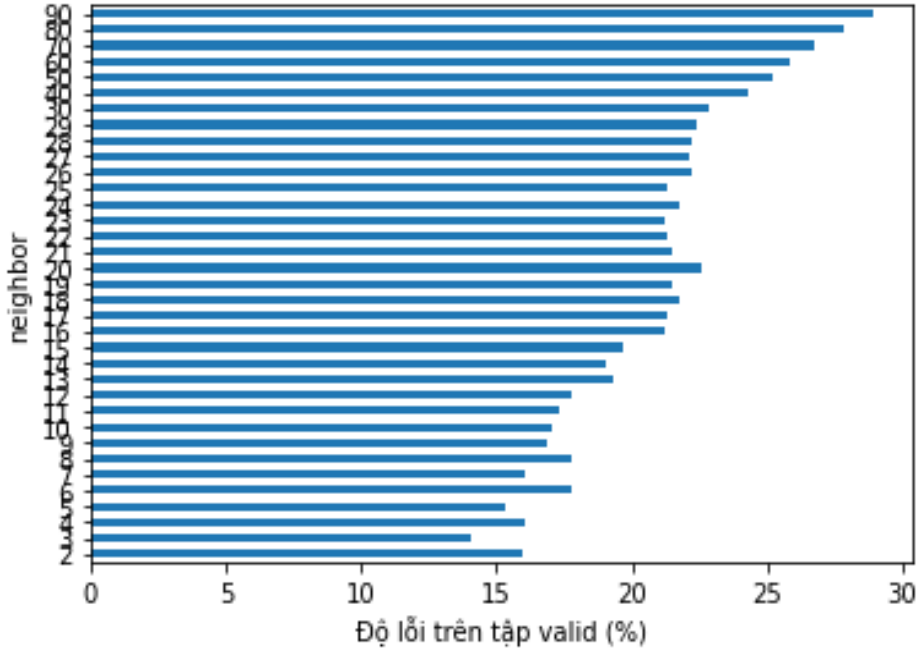
•Mô hình fit khá ổn trên tập dữ liệu.

Mô hình K-neighbor Classifier

- Thử nghiệm nhiều giá trị của siêu tham số n_neighbors để tìm được độ lỗi trên tập validation tối ưu.

Độ lỗi trên các tập dữ liệu





Độ lỗi nhỏ nhất trên tập valid là 14.06%, neighbor tốt nhất là 3

Huấn luyện lại mô hình với cả tập train và tập valid

Với neighbor = 3:
- Độ lỗi trên tập train + valid: 5.39%
- Độ lỗi trên tập test: 13.89%

Nhận xét:
•Độ lỗi trên cả hai tập chênh lệch nhiều .
•Độ lỗi trên tập test cao (trên 10%)
•Dựa vào hai biểu đồ độ lỗi trên, ta thấy tham số n_neighbors không tỉ lệ thuận với độ lỗi trên cả hai tập.Trên miền giá trị n_neighbors đã xét, giá trị độ lỗi có lúc tăng , có lúc giảm xen kẽ nhau.

Mô hình tốt nhất

=> Dựa vào độ lỗi trên tập test trên ta thấy mô hình MLP Classifier hoạt động hiệu quả hơn hai mô hình còn lại => dự đoán chất lượng không khí tốt nhất

7

Quá trình làm đồ án

Những khó khăn đã gặp phải

- **Quang**: lựa chọn mô hình phù hợp với dữ liệu đã dùng, tốn thời gian trong việc tìm kiếm nguồn cung dữ liệu (một số webiste yêu cầu trả phí hoặc gửi email xin xác nhận mục đích sử dụng để truy cập dữ liệu).

- **Sơn**:
  - Tìm kiếm các ý tưởng.
  - Có ý tưởng nhưng không thể tìm được dữ liệu.
  - Tìm được các API phù hợp nhưng phải trả phí.

Những kỹ năng đã học được

- **Quang**: qua đồ án này em có thêm kinh nghiệm để chọn lọc và thu thập dữ liệu, hiểu thêm nhiều thuật toán phân lớp hữu ích và cẩn thận hơn trong cách trình bày.
- **Sơn**:
  - Nhiều trang cung cấp api miễn phí thú vị
  - Dùng các thuật toán phân lớp để mô hình hóa dữ liệu
  - Bình tĩnh hơn khi gặp các vấn đề không thể giải quyết ngay

Nếu có thêm thời gian

Nhóm sẽ tìm và thu thập dữ liệu đa dạng, phức tạp hơn, có thể thử nghiệm thu thập dữ liệu bằng cách parse HTML và mô hình hóa dữ liệu với nhiều thuật toán khác.

8 Tài liệu tham khảo

Chủ yếu từ các trang web:

- https://openweathermap.org/api/air-pollution
- https://scikit-learn.org/
- https://stackoverflow.com/
- https://www.w3schools.com/

Video bài giảng + demo của môn học + slide bài giảng
Homework:1, 2, 3

Cảm ơn thầy đã lắng nghe