|  |
| --- |
| TRƯỜNG ĐẠI HỌC THỦY LỢI  **KHOA CÔNG NGHỆ THÔNG TIN**  A blue and white logo  Description automatically generated with low confidence  **BÀI TẬP LỚN**  HỌC PHẦN: HỌC MÁY  **ĐỀ TÀI: DỰ ĐOÁN MỨC TIÊU THỤ ĐIỆN CỦA**  **THÀNH PHỐ TETOUAN**      **Giáo viên hướng dẫn**: Nguyễn Huy Đức  **Nhóm sinh viên thực hiện**:  1. Đỗ Trung Kiên, Lớp 62TH-NB  2. Trần Thanh Thảo, Lớp 62TH-NB  3. Đặng Văn Dương, Lớp 62TH-NB  **Hà Nội, 10/2022** |

**Phần 1: Tổng quan**

***1.Mô tả bài toán***

* **Tên bài toán**: Dự đoán mức tiêu thụ điện của thành phố Tetouan
* **Input**:
* Temperature: Weather Temperature of Tetouan city.
* Humidity: Weather Humidity of Tetouan city.
* Wind Speed of Tetouan city.
* general diffuse flows
* diffuse flows
* **Ouput:**
* power consumption of zone 1 of Tetouan city.
* power consumption of zone 2 of Tetouan city.
* power consumption of zone 3 of Tetouan city.
* **Tóm tắt công việc thực hiện của bài toán**: Tập dữ liệu này liên quan đến mức tiêu thụ điện năng của ba mạng lưới phân phối khác nhau của thành phố Tetouan nằm ở phía bắc Maroc
* Đọc dữ liệu từ file Tetuan City power consumption.csv
* Dùng Kfold để chia tập dữ liệu thành 70% huấn luyện, 30% kiểm tra
* Dùng hàm LinearRegression trong thư viện sklearn để huấn luyện mô hình
* Dùng explained\_variance\_score để đánh giá
* So sánh kết quả train và dữ liệu test
* Tính kết quả trung bình và độ lệch

2***. Phương pháp học máy***

* ***Phương pháp Hồi quy tuyến tính***
* Phân tích hồi quy tuyến tính là tìm sự liên hệ giữa 2 biến số liên tục: biến độc lập (biến dự đoán) x với biến phụ thuộc (biến kết cục) y. Sau đó vẽ một đường thẳng hồi quy và từ phương trình đường thẳng này ta có thể dự đoán được biến y.

y = βx + α

* Dữ liệu của bài toán có thể nằm trong không gian nhiều hơn 2 chiều. Khi đó, output y chính là sự kết hợp các input xi theo một tỷ lệ nào đó. Tỷ lệ này là các hệ số wi gọi là trọng số của mô hình. Các giá trị xi có thể được viết thành vector X, các trọng số wi có thể được viết thành vector W. Việc tối ưu mô hình Linear Regression là tìm ra vector W sao cho từ input X ta có thể tính ra được output y của bài toán.

y  = XW

**Sai số dự đoán** e = y - y được gọi là sai số dự đoán. Giá trị này sẽ được tối ưu sao cho gần 0 nhất.

* Hàm mất mát: đánh giá độ lỗi, độ sai của mô hình. Kết quả của hàm này có giá trị càng lớn thì mô hình của chúng ta càng sai.

L w = 12Ni=1N(yi-xiTW)2=12N||y-XTw||22

* Tìm nghiệm cho bài toán: Giải phương trình đạo hàm của hàm mất mát bằng 0:

                 ∇L(w)∇w=0      ⬄   XXTw = Xy    ⬄      w=(XXT)-1Xy

* ***Phương pháp k-fold cross validation***
  + Cross validation là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm.Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá.
  + Kỹ thuật này thường bao gồm các bước như sau:

1. Xáo trộn dataset một cách ngẫu nhiên.
2. Chia dataset thành k nhóm
3. Với mỗi nhóm:

* Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
* Các nhóm còn lại được sử dụng để huấn luyện mô hình
* Huấn luyện mô hình
* Đánh giá và sau đó hủy mô hình

1. Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá

* Kết quả tổng hợp thường là trung bình của các lần đánh giá.

Cấu hình k:k là một số nguyên, thường là 5 hoặc 10

**Phần 2: Thực nghiệm**

***1.Mô tả tập dữ liệu của bài toán***

* Dữ liệu gồm 8 thông tin
* Có 52417 vector dữ liệu
* Mô tả ma trận dữ liệu : Đưa ra dữ liệu về mức tiêu thụ điện năng của ba mạng lưới phân phối khác nhau của thành phố Tetouan
* Mô tả ma trận dữ liệu (X) gồm 5 thông tin:

• Temperature: nhiệt độ.

• Humidity: Độ ẩm.

• Wind Speed: Tốc độ gió

• general diffuse flows : dòng khuếch tán chung dòng

• diffuse flows: khuếch tán

* Nhãn lớp (Y) gồm là 3 thông tin:
* power consumption of zone 1 of Tetouan city: mức tiêu thụ điện năng của mạng lưới 1
* power consumption of zone 2 of Tetouan city: mức tiêu thụ điện năng của mạng lưới 2
* power consumption of zone 3 of Tetouan city: mức tiêu thụ điện năng của mạng lưới 3

***2. Phân tích kết quả của chương trình***

- Tỷ lệ dự đoán đúng:

Lần 1: 0.03560621

Lần 2: 0.12473938

Lần 3: 0.05966984

- Tỉ lệ trung bình: 0.733

- Tỷ lệ dự đoán sai: 0.0376

**Kết luận**

* Tóm lược các nội dung chính mà bài tập lớn làm được
* Tìm ra tỷ lệ dự đoán đúng, tỉ lệ dự đoán sai.
* Cho ra kết quả dự đoán dữ liệu mức tiêu thụ điện năng của ba mạng lưới phân phối khác nhau của thành phố Tetouan

**Tài liệu tham khảo**

* Các tài liệu được tham khảo trong báo cáo:
* Bài giảng của giáo viên
* Kho lưu trữ UCI: https://archive.ics.uci.edu/ml/index.php