TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: DỰ ĐOÁN CHI PHÍ Y TẾ**

Giáo viên hướng dẫn: Nguyễn Huy Đức

Nhóm 14 thực hiện:

1. Trần Trung Lâm, Lớp 62 TH-VA

2. Phạm Việt Hùng, Lớp 62 TH-VA

3. Hà Xuân Trường, Lớp 62 TH-VA

**Hà Nội, năm 2022**



**Phần 1: Tổng quan**

***1.Mô tả bài toán***

* Tên bài toán: Dự đoán chi phí y tế
* Input:

+ Vector X gồm 7 trường: (498 vector)

1. tuổi
2. giới tính
3. bmi
4. số con cái
5. hút thuốc
6. nơi ở
7. chi phí

* Output:
  + 1. Đưa ra chi phí y tế dự đoán.
    2. W, W0
    3. Giá dự đoán, giá thực tế, chênh lệch
    4. Điểm đánh giá mức độ chênh lệch
* Tóm tắt công việc thực hiện của bài toán :
* Hồi quy tuyến tính

+ Đọc dữ liệu từ tập dữ liệu insurance.csv

+ Chuyển các cột ‘Gioi\_Tinh,’ Hut\_thuoc’,’ Vung’ từ chuỗi thành số dùng hàm LabelEncoder

+ Chia tập dữ liệu 70% dùng để huấn luyện, 30% dùng để kiểm tra.

+Dùng hàm LinearRegression để huấn luyện mô hình

+In ra kết quả.

- Phương pháp KFold:

+ Đọc dữ liệu từ tập dữ liệu insurance.csv

+ Chuyển các cột ‘Gioi\_Tinh,’ Hut\_thuoc’,’ Vung’ từ chuỗi thành số dùng hàm LabelEncoder

+ Chia tập dữ liệu 70% dùng để huấn luyện, 30% dùng để kiểm tra.

+Chia tập huấn luyện thành k -1 phần, 1 phần còn lại dùng để kiểm tra

+ Xây dựng hàm để tính trung bình lỗi giữa y dự đoán và y thực tế.

+ Lấy ra mô hình tốt nhất

+In ra kết quả

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 qui 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  chính là sự kết hợp các input  theo một tỷ lệ nào đó. Tỷ lệ này là các hệ số  gọi là trọng số của mô hình. Các giá trị  có thể được viết thành vector X, các trọng số  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  của bài toán.

**Sai số dự đoán** e = 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.
* Tìm nghiệm cho bài toán: Giải phương trình đạo hàm của hàm mất mát bằng 0:

⬄ X = Xy ⬄

* Lý thuyết k-fold cross validation:
  + Trong xác thực chéo, tập huấn luyện được chia thành k tập con có kích thước gần bằng nhau và không giao nhau. Tại mỗi lần thử, một trong k tập con đó được lấy ra làm tập xác thực, k − 1 tập con còn lại được coi là tập huấn luyện. Như vậy, với mỗi bộ tham số mô hình, ta có k mô hình khác nhau. Sai số huấn luyện và sai số xác thực được tính là trung bình cộng của các giá trị tương ứng trong k mô hình đó. Cách làm này có tên gọi là xác thực chéo k-fold (k-fold cross validation).
  + Kỹ thuật này bao gồm các bước sau:
    - B1: Xáo trộn dataset một cách ngẫu nhiên
    - B2: Chia dataset thành k nhóm
    - B3: Với mỗi nhóm:
      * Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
      * Các nhóm còn lại sẽ được sử dụng để huấn luyện mô hình
      * Huấn luyện mô hình
      * Đánh giá và hủy mô hình
* B4: Tổng hợp hiệu quả mô hình dựa tren kết quả đánh giá

**Phần 2: Thực nghiệm**

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

* Vector X gồm 7 trường: tuổi, giới tính,bmi,số con cái,hút thuốc,nơi ở, chi phí. có 498 mẫu dữ liệu .
* Nhãn lớp Y= [chi phí]

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

* Tỷ lệ dự đoán đúng:

+ Hồi quy tuyến tính: 74% đến 80%

+ KFold: 76% đến 80%

* Tỷ lệ dự đoán sai:

+ Hồi quy tuyến tính: 20% đến 26%

+ KFold: 20% đến 24%

**Kết luận**

* Tóm lược các nội dung chính mà bài tập lớn làm được:
* Nắm được phương pháp Hồi quy tuyến tính và phương pháp KFold để xử lý bài toán dự đoán chi phí bảo hiểm y tế.
* Thực hiện bài toán dự đoán trên 2 thuật toán hồi quy tuyến tính (linear\_regression) và xác thực chéo (k-fold cross validation).
* Đưa ra các bước cơ bản nhất để xây dựng 1 mô hình học máy dựa trên các thuật toán.

**Tài liệu tham khảo**

<https://priscilla-vanny.medium.com/linear-regression-with-k-fold-cross-validation-in-python-predict-the-health-insurance-cost-7badf90fc1b8>

<https://archive.ics.uci.edu/ml/index.php>

<https://machinelearningcoban.com/2016/12/28/linearregression/>