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: PHÂN TÍCH NHÂN SỰ:**

**DỮ LIỆU THĂNG TIẾN CỦA NHÂN VIÊN**

Giáo viên hướng dẫn:   
 Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện (Nhóm 10):

1. Đàm ViếtVương, lớp 62TH1

2. Nguyễn Mạnh Tiến, lớp 62TH1

3. Đỗ Văn Thông, lớp 62TH1

**Hà Nội, năm 2022**

**Phần 1: Tổng quan**

***1.Mô tả bài toán***

* Tên bài toán: Phân tích dữ liệu điểm đánh giá của nhân viên từ đó dự đoán khả năng được tăng lương.
* Input: Dữ liệu nhân sự về giới tính, tuổi tác, các điểm đánh giá.
* Output:  Phần trăm khả năng được tăng lương của nhân viên.
* Tóm tắt công việc:
* Đọc tập dữ liệu của bài toán
* Chia dữ liệu thành 2 phần: 70% cho tập train, 30% cho tập test
* Sử dụng phương pháp k-fold cross validation chia dữ liệu của tập train thành k phần để tìm mô hình tối ưu nhất
* Huấn luyện mô hình và dự đoán nhãn

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  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 ⬄

* 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***

* Mỗi mẫu Vector gồm 7 thuộc tính: : Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, PercentSalaryHike. Có 1470 mẫu vector dữ liệu.
* Vector X = [[Age] , [DailyRate] , [Education] , [HourlyRate] , [JobLevel] , [MonthlyRate]].
* Age: Tuổi của nhân viên
* DailyRate: Điểm đánh giá hàng ngày
* Education: Trình độ học vấn đánh giá theo các cấp độ từ 1 đến 5  
  1 : Dưới đại học  
  2 : Đại học  
  3 : Cử nhân  
  4 :Thạc sĩ  
  5 : Tiến sĩ
* HourlyRate: Điểm đánh giá hàng giờ
* JobLevel: Cấp bậc công việc
* MonthlyRate: Điểm đánh giá hàng tháng
* Nhãn lớp Y = [PercentSalaryHike].

PercentSalaryHike: Khả năng được tăng lương (%)

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

Coi những mẫu độ chênh lệch <=1 là những mẫu dự đoán đúng, ta thu được:

* Số mẫu dự đoán đúng: 69/441, Tỷ lệ dự đoán đúng: 16%
* Số mẫu dự đoán sai: 372/441, Tỷ lệ dự đoán sai: 84%
* Giá trị chênh lệch trung bình: 3.02
* Hiệu suất của mô hình hồi quy tuyến tính (Coefficient of determination): -0.01

**Kết luận**

* Xử lý thông tin của bộ dữ liệu ban đầu chọn ra những thông tin liên quan đến việc dự đoán khả năng tăng lương của nhân viên
* Sử dụng phương pháp hồi quy tuyến tính và kĩ thuật k fold cross validation dự đoán được khả năng tăng lương của nhân viên với mức chênh lệch trung bình là 3.

**Tài liệu tham khảo**

* Các tài liệu được tham khảo trong báo cáo
* <https://machinelearningcoban.com/2016/12/28/linearregression/>
* [https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?fbclid=IwAR1p4VsXA67AFr4t8_k2-R4cmXxjwEmEZLFDUkV_ebJ06JFTD3qGQLpbRIU)