BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



KHOA CÔNG NGHỆ THÔNG TIN

NHẬP MÔN DỮ LIỆU LỚN

**BÁO CÁO BÀI TẬP MÔN HỌC NỘI DUNG: LAB 03 – MACHINE LEARNING ON SPARK**

**Giảng viên hướng dẫn** : Huỳnh Lâm Hải Đăng

**Lớp** : CQ2022/21

**Sinh viên thực hiện** : Trương Tiến Anh- 22120017 Nguyễn Minh Cường- 22120043 Đinh Viết Lợi- 22120188 Nguyễn Trần Lợi-22120190

*Hồ Chí Minh, ngày 6 tháng 5 năm 2025*

# MỤC LỤC

[MỤC LỤC 1](#_Toc197443066)

[PHẦN 1: CLASSIFICATION WITH LOGISTIC REGRESSION 2](#_Toc197443067)

[I. Structured API Implementation (High Level) 2](#_Toc197443068)

[1. Ý tưởng thực hiện 2](#_Toc197443069)

[2. Mô tả chi tiết 2](#_Toc197443070)

[II. MLlib RDD-Based Implementation 7](#_Toc197443071)

[1. Ý tưởng thực hiện 7](#_Toc197443072)

[2. Mô tả chi tiết 7](#_Toc197443073)

[III. Low-Level Operations 8](#_Toc197443074)

[1. Ý tưởng thực hiện 8](#_Toc197443075)

[2. Mô tả chi tiết 8](#_Toc197443076)

[IV. So sánh các phương pháp giải quyết 8](#_Toc197443077)

[PHẦN 2: REGRESSION WITH DECISION TREES 9](#_Toc197443078)

[I. Structured API Implementation (High Level) 9](#_Toc197443079)

[1. Ý tưởng thực hiện 9](#_Toc197443080)

[2. Mô tả chi tiết 9](#_Toc197443081)

[II. MLlib RDD-Based Implementation 10](#_Toc197443082)

[1. Ý tưởng thực hiện 10](#_Toc197443083)

[2. Mô tả chi tiết 10](#_Toc197443084)

[III. Low-Level Operations 11](#_Toc197443085)

[1. Ý tưởng thực hiện 11](#_Toc197443086)

[2. Mô tả chi tiết 11](#_Toc197443087)

[IV. So sánh các phương pháp giải quyết 11](#_Toc197443088)

[PHẦN 3: BÁO CÁO NHÓM 11](#_Toc197443089)

# PHẦN 1: CLASSIFICATION WITH LOGISTIC REGRESSION

## Structured API Implementation (High Level)

### **Ý tưởng thực hiện**

* + - Structured API cung cấp lượng lớn các API giúp xử lý dữ liệu có scheme rõ ràng và hầu như không yêu cầu viết rất ít các hàm bổ trợ hay xử lý thủ công.
    - Tập dữ liệu được đưa ra cung cấp các giao dịch bằng thẻ tín dụng tại Châu Âu. Do yêu cầu về bảo mật riêng tư nên tên thuộc tính và giá trị tất cả các thuộc tính ngoại trừ `Thời gian giao dịch` và `Số tiền giao dịch` đều đã được chuẩn hoá và ẩn đi.
    - Một vấn đề lớn của tập dữ liệu này là hiện tượng mất cân bằng dữ liệu cực nặng khi mà chỉ có 492 trên 284,807 giao dịch là giao dịch lừa đảo.
    - Bài làm sẽ sử dụng Structured API với mục đích sau cùng là xây dựng một mô hình nhằm xác định xem giao dịch nào là lừa đảo. Quy trình sẽ bao gồm: khám phá dữ liệu, tiền xử lý dữ liệu cơ bản, phân chia tập huấn luyện, huấn luyện mô hình, cải thiện mô hình bằng phương pháp Undersampling.
    - Việc ghi nhận các giao dịch lừa đảo là rất quan trọng, hơn rất nhiều so với các giao dịch không phải lừa đảo nên mô hình cũng sẽ được xây dựng theo hướng tối ưu nhất các dự đoán giao dịch là lừa đảo.

### **Mô tả chi tiết**

* + - Khởi tạo Session và đọc dữ liệu đầu vào

*A screenshot of a computer

AI-generated content may be incorrect.*

*Hình 1: Tập dữ liệu*

* + - Kiểm tra missing values

A screenshot of a computer

AI-generated content may be incorrect.

*Hình 2: Kiểm tra missing values*

* + - Chuẩn hoá dữ liệu: Thực hiện chuẩn hoá Z-score đối với 2 dữ liệu chưa được chuẩn hoá là `time` và `Amount`, ta gom 2 thuộc tính lại thành 1 vector để chuẩn hoá sau đó tách ra lại.

A computer screen with white text

AI-generated content may be incorrect.

*Hình 3: Chuẩn hoá dữ liệu*

* + - Xoá dữ liệu trùng lắp: dữ liệu bị trùng sẽ ảnh hưởng đến chất lượng của mô hình

*A screen shot of a computer

AI-generated content may be incorrect.*

*Hình 4: Xoá dữ liệu trùng lặp*

* + - Phân chia tập dữ liệu: tập dữ liệu không những chỉ phải được phân chia với tỉ lệ train-test phù hợp mà còn phải đảm bảo số lượng nhãn được phân chia phù hợp với vấn đề mất cân bằng dữ liệu trầm trọng.

*A screenshot of a computer program

AI-generated content may be incorrect. Hình 5: Kết quả phân chia train-test*

* + - Huấn luyện mô hình cơ bản: sau khi tiền xử lý dữ liệu cơ bản, ta có thể tiến hành học từ tập huấn luyện nhanh chóng khi structured API hỗ trợ các API cho mô hình logistic regression sau khi assemble các thuộc tính thành vector.

*A screen shot of a computer code

AI-generated content may be incorrect.*

*Hình 6: Huấn luyện mô hình*

* + - Đánh giá mô hình trên tập huấn luyện: Đánh giá mô hình bằng các chỉ số Accuracy, AUC, Precision, Recall và F1-score. Mô hình hiện tại cho thấy hiệu suất đối với lớp 0 cực cao trong khi với lớp 1 thì chỉ đạt mức trung bình, điều này là do việc mất cân bằng dự liệu.

*A screenshot of a computer

AI-generated content may be incorrect.*

*Hình 7: Hiệu suất mô hình trên tập huấn luyện*

* + - Dự đoán và kiểm tra trên tập test: Áp dụng mô hình vừa học được ở trên cho dữ liệu của tập test để bổ sung thêm cột dự đoán, tính các chỉ số tp,tn,fp,fn để tính các thông số hiệu suất. Tuy nhiên kết quả chưa thể chấp nhận được đối với nhãn 1 nên ta sẽ tiếp tục cải thiện hiệu suất mô hình.

*A screen shot of a computer

AI-generated content may be incorrect.*

*Hình 8: Chỉ số trên tập test*

* + - Undersampling: áp dụng phương pháp undersampling để giảm kích thước của nhãn 0 xuống để giúp mô hình học tốt hơn, có thể phân biệt rõ ràng nhãn 0 và nhãn 1. Lưu ý chỉ oversampling sau khi split tập train-test. Ta lấy tỉ lệ 1:80 thay vì 1:577 như tập dữ liệu gốc.

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 9: Undersampling

* + - Vẽ biểu đồ tương quan để thấy kết quả của undersampling ảnh hưởng thế nào đến mô hình

*A graph of numbers and symbols

AI-generated content may be incorrect.*

*Hình 10: Ma trận tương quan sau Undersampling*

* + - Huấn luyện lại mô hình trên tập Undersampling: các thông số đánh giá đối với nhãn 1 có sự cải thiện đáng kể.

*A computer screen shot of a number

AI-generated content may be incorrect.*

*Hình 11: Hiệu suất mô hình trên tập Undersampling*

* + - Dự đoán và kiểm tra tập test sau khi huấn luyện lại mô hình: Chỉ số Recall của mô hình đã có sự cải thiện đáng kể, giúp việc dự đoán một giao dịch là lừa đảo trở nên chính xác hơn.

*A screen shot of a computer

AI-generated content may be incorrect.*

*Hình 12: Chỉ số trên tập test*

## MLlib RDD-Based Implementation

### **Ý tưởng thực hiện**

* + - Report date thực hiện các báo cáo “in the last 7 days”, tức là ngày thứ hai mỗi tuần sẽ thực hiện bản báo cáo của 7 ngày tuần trước- từ sáng thứ 2 tuần trước tới tối chủ nhật tuần trước.

### **Mô tả chi tiết**

* + - Tạo Park Session và import các thư viện cần thiết.

## Low-Level Operations

### **Ý tưởng thực hiện**

* + - Xem các điểm tạo thành tứ giác bao lồi.

### **Mô tả chi tiết**

* + - Import các thư viện cần thiết và tạo Spark Session.

## So sánh các phương pháp giải quyết

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| Structured API-High Level | Hiệu suất cao nhờ Catalyst Optimizer & Tungsten Engine.  Dễ viết, dễ đọc khi cú pháp tương tự SQL.  Hỗ trợ nhiều API có sẵn, tránh xử lý thủ công.  Dễ dàng build các pipleline trong machine learning | Chỉ hỗ trợ xử lý dữ liệu có cấu trúc, có scheme rõ ràng. |
| MLLib RDD-Based |  |  |
| Low-Level Operations |  |  |

# PHẦN 2: REGRESSION WITH DECISION TREES

## Structured API Implementation (High Level)

### **Ý tưởng thực hiện**

### **Mô tả chi tiết**

* + - Mapper (SlidingWindowMapper):

## MLlib RDD-Based Implementation

### **Ý tưởng thực hiện**

* + - Report date thực hiện các báo cáo “in the last 7 days”, tức là ngày thứ hai mỗi tuần sẽ thực hiện bản báo cáo của 7 ngày tuần trước- từ sáng thứ 2 tuần trước tới tối chủ nhật tuần trước.

### **Mô tả chi tiết**

* + - Tạo Park Session và import các thư viện cần thiết.

## Low-Level Operations

### **Ý tưởng thực hiện**

* + - Xem các điểm tạo thành tứ giác bao lồi.

### **Mô tả chi tiết**

* + - Import các thư viện cần thiết và tạo Spark Session.

## So sánh các phương pháp giải quyết

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| Structured API-High Level |  |  |
| MLLib RDD-Based |  |  |
| Low-Level Operations |  |  |

# PHẦN 3: BÁO CÁO NHÓM

* + Phương thức trao đổi và làm việc:
    - Vì khối lượng công việc lớn đồng thời để đảm bảo chất lượng, nhóm phân chia thành 2 đội nhóm chịu trách nhiệm giải quyết hai vấn đề lớn.
    - Nhóm tổ chức phiên họp nhóm sau một khoảng thời gian làm việc cá nhân để theo dõi sát tình hình giải quyết bài tập.
    - Phiên họp nhằm kiểm tra kết quả, ý tưởng giải quyết cho từng vấn đề.
    - Giao công việc cho từng thành viên trên workspace của nhóm sau khi thống nhất kết quả.
  + Bảng phân công công việc:

|  |  |  |
| --- | --- | --- |
| Thành viên | Công việc | Lưu ý |
| Trương Tiến  Anh | Thực hiện các phương thức giải quyết bài tập đối với vấn đề Regression with Decision Trees.  Khảo sát chất lượng các file, tổ chức thư mục và chú thích tại từng file. | Cả hai thành viên giải quyết vấn đề cùng nhau, chia sẻ các phương pháp cho từng cách tiếp cận và cân nhắc tính khả dụng ở các cách tiếp cận khác. |
| Nguyễn Minh  Cường |
| Đinh Viết Lợi | Thực hiện các phương thức giải quyết bài tập đối với vấn đề Classification with Logistic Regression.  Viết báo cáo tổng hợp bài tập nhóm. |  |
| Nguyễn Trần  Lợi |