Lab
Lập trình Spark với Python

1. Mô tả

Trong bài lab này, sinh viên thực hành viết mã nguồn PySpark để thực thi một số tác vụ khai thác dữ liệu.

2. Yêu cầu

Sinh viên hoàn thiện các bài tập nhỏ sau.

2.1. Khai thác mẫu phổ biến và luật kết hợp (20%)

Trong bài tập này, bạn sẽ làm việc trên các tập tin orders.csv và products.csv (xem tập tin đính kèm) biểu diễn dữ liệu mua hàng được lấy mẫu từ hơn ba triệu đơn hàng tạp hóa của 200,000 người dùng Instacart users.

- Tập tin orders.csv chứa các giao dịch đã thực hiện. Mỗi giao dịch được thể hiện bằng một hay nhiều dòng, thông tin trên mỗi dòng bao gồm: mã giao dịch (order_id), mã sản phẩm (product_id), thứ tự bỏ vào giỏ (add_to_cart_order), và sản phẩm này có được mua lại hay không (reordered). Ta chỉ quan tâm hai trường đầu tiên.
- Tập tin products.csv chứa các sản phẩm bày bán trong cửa hàng. Mỗi dòng trình bày thông tin của một sản phẩm, bao gồm: mã sản phẩm (product_id), tên sản phẩm (product_name), và thông tin khác (aisle_id và department_id). Ta chỉ quan tâm hai trường đầu tiên.

Bạn cần áp dụng giải thuật khai thác mẫu phổ biến và luật kết hợp của MLlib lên dữ liệu được cho ở trên. Để hoàn tất yêu cầu của câu hỏi, các bước cần thực hiện là

1) Đọc hai tập tin vào PySpark và sử dụng các phương thức tiền xử lý dữ liệu phù hợp để đưa dữ liệu về một DataFrame gồm hai cột theo đúng thứ tự là mã giao dịch (order_id) và danh sách sản phẩm (thể hiện bằng tên, không phải mã sản phẩm.

id	products
1	['Beef', 'Chicken', 'Milk']

2) Áp dụng giải thuật khai thác mẫu phổ biến và luật kết hợp trong gói pyspark.ml.fpm. Thử nghiệm với một số bộ giá trị ngưỡng support và confidence.

3) Bạn có nhận thấy vấn đề gì về hình thức của các luật được tìm thấy hay không (ta không cần quan tâm ngữ nghĩa của dữ liệu). Nếu có, hãy khắc phục điều này.

2.2. Bài 2 – Phân lớp (20%)

Trong bài tập này, bạn sẽ làm việc trên tập tin mushroom.csv biểu diễn dữ liệu các loài nấm. Dữ liệu có 8124 mẫu, trong đó mỗi mẫu được thể hiện bằng 22 thuộc tính và phân loại thành "edible" (e) hoặc "poisonous" (p).

Bạn cần thực nghiệm các giải thuật phân lớp của MLlib trên dữ liệu được cho. Để hoàn tất yêu cầu của câu hỏi, các bước cần thực hiện là

1) Chia dữ liệu thành tập huấn luyện và tập kiểm thử theo tỉ lệ 80:20
2) Xây dựng mô hình decision tree trên tập huấn luyện
3) Xây dựng mô hình random forest trên tập huấn luyện
4) Đánh giá hai mô hình trên tập kiểm thử.
5) Sử dụng pipeline để thiết lập các bước trên thành một luồng xử lý duy nhất.

2.3. Bài 3 – Gom cụm (20%)

Trong bài tập này, bạn sẽ làm việc trên tập dữ liệu plants về sự phân bố của một số loài thực vật ở khu vực Mỹ và Canada (tải dữ liệu từ http://archive.ics.uci.edu/ml/datasets/Plants). Lưu ý: có vùng phân bố không được liệt kê trong tập tin mô tả dữ liệu, cần đối chiếu với tập tin dữ liệu để xác định đúng danh sách vùng phân bố.

Bạn cần thực nghiệm giải thuật gom cụm k-Means của MLlib trên dữ liệu được cho. Để hoàn tất yêu cầu của câu hỏi, các bước cần thực hiện là

1) Đọc các tập tin cần thiết vào PySpark và áp dụng kỹ thuật tiền xử lý dữ liệu phù hợp để chuyển đổi dữ liệu về dạng vectơ nhị phân. Như vậy ta sẽ có một DataFrame bao gồm:
    - Cột đầu tiên thể hiện tên loài thực vật và các cột tiếp theo biểu diễn vùng địa lý.
    - Mỗi dòng thể hiện thông tin phân bố địa lý của nột loài thực vật – nếu loài thực vật có tại một vùng địa lý thì ô tương ứng mang giá trị 1, ngược lại mang giá trị 0.

    Lưu dữ liệu sau khi tiền xử lý thành tập tin plants.csv. Đính kèm khi nộp.
2) Chương trình thực hiện gom cụm các loài thực vật theo vùng địa lý bằng giải thuật k-means và đánh giá kết quả gom cụm bằng ClusteringEvaluator. Thử nghiệm với một số giá trị k.

3. Thang điểm

No.	Criteria	Scores
1	Hoàn thành mỗi bài với tỉ lệ đã được ghi bên cạnh	60%
2	Quay video chứng minh quá trình chạy và thực thi thành công. Nếu không có video thì không xem xét là hoàn thành các bài tập.	10%
3	Báo cáo đầy đủ và chi tiết quá trình thực hiện, mã nguồn. Trình bày rõ ràng, bố cục hợp lý. Tổ chức các tập tin thành từng thư mục thể hiện ý nghĩa của mỗi nhóm như thư mục mã nguồn, báo cáo, test, v.v…	30%
Total		100%

4. Lưu ý

- Bài lab được thực hiện dưới dạng nhóm.
- Hạn chót theo thông tin trên Moodle.
- Nếu thực hiện mã nguồn trên Google Colab cần cấp quyền thao tác cho các giáo viên của môn học và tự đảm bảo không thao tác gì khác sau hạn nộp.
- Ngoài các yêu cầu về nội dung, báo cáo cần có các thông tin cơ bản sau:
  - Thông tin về các thành viên
  - Kế hoạch và phân công
  - Tự đánh giá mức độ hoàn thành của mỗi thành viên cho các công việc phụ trách.
  - Tài liệu tham khảo (nếu có)
- Đạo văn, gian lận trong quá trình làm bài sẽ 0 điểm môn học.