Ton Duc Thang University
Faculty of Information Technology

ĐỒ ÁN CUỐI KỲ

Môn: Xử lý dữ liệu lớn

Hạn nộp bài: 19/12/2022

I. Hình thức

- Đồ án cuối kỳ được thực hiện theo nhóm **02 – 03** sinh viên.
- Sinh viên thực hiện đồ án không theo nhóm sẽ bị trừ điểm.
- Nhóm sinh viên thực hiện các yêu cầu và nộp bài theo hướng dẫn chi tiết bên dưới.

II. Yêu cầu

Cho các tập dữ liệu tại thư mục datasets, sinh viên thực hiện các yêu cầu sau.

Tệp dữ liệu	Mô tả
oxford_pet3_train.csv	Dữ liệu hình ảnh và chủng loại thú cưng chọn ra từ tập Oxford Pet III[1]. 500 dòng dữ liệu. Mỗi dòng chứa 49154 số nguyên • Số thứ nhất: chỉ số dòng • Số thứ hai: chủng loại thú cưng • Còn lại: 128 * 128 * 3 số nguyên là ảnh thú cưng dạng vector 1 chiều.
oxford_pet3_test.csv	Tương tự **oxford_pet3_train.csv**
ratings2k.csv	Dữ liệu đánh giá sản phẩm Dòng 1 là header • **index**: chỉ số dòng • **user**: mã người dùng

[1] https://www.robots.ox.ac.uk/~vgg/data/pets/

	• **item**: mã món hàng • **rating**: đánh giá (0.0-5.0) 2365 dòng tiếp theo là dữ liệu tương ứng
stockHVN2022.csv	Dữ liệu mã chứng khoán HVN trên sàn HOSE trong năm 2022 (đến ngày 18/11). Dòng 1 là header: • **Ngay**: ngày ghi nhận • **HVN**: giá đóng cửa

a) Câu 1 (2.0 điểm): Giảm số chiều với SVD

i. Hiển thị

Sinh viên đọc tệp **oxford_pet3_train.csv** sử dụng **DataFrame** của **pyspark.sql** và sử dụng thư viện **matplotlib.pyplot** để vẽ ra biểu đồ 3 x 5 ô hiển thị 15 bức ảnh đầu tiên.

- Chuyển đổi vector ảnh 49152 chiều thành ma trận 128 x 128 x 3
- Hiển thị ảnh với hàm **imshow**()
- Với từng bức ảnh trong lưới 3 x 5, hiển thị **title** là label (chủng loại) của thú cưng.

ii. Giảm số chiều tập train

Sinh viên chọn ra 01 số chiều **r** ***lớn nhất có thể*** trong đoạn [1, 49152]. Với giá trị **r**, sinh viên sử dụng thuật toán **SVD** để giảm số chiều cho toàn bộ vector trong tập **oxford_pet3_train.csv**. Sau đó lưu lại kết quả thành 01 tệp csv có cấu trúc tương tự **oxford_pet3_train.csv** (lưu ý tên tệp có thêm số **r** để dễ phân biệt).

iii. Giảm số chiều tập test

Tương tự với mục ii, sinh viên thực hiện giảm số chiều cho tập **oxford_pet3_test.csv** và lưu kết quả thành tệp csv tương tự, **trong đó lưu ý giá trị r cho tập train và test phải bằng nhau.**

b) Câu 2 (2.0 điểm): Khuyến nghị sản phẩm với Collaborative Filtering

Sinh viên sử dụng tập **ratings2k.csv** cho câu này để tạo thành một **DataFrame** (**pyspark.sql**) trong đó

- Mỗi dòng ứng với một **user** (74 users)
- Mỗi cột ứng với một **item** (467 items)
- Các dòng được xếp tăng dần theo người dùng.

Sinh viên sử dụng thông tin của 70 người dùng đầu tiên, bằng phương pháp **Collaborative Filtering**, tính ra tất cả ratings cho 4 người dùng còn lại trong đó so sánh độ tương đồng bằng **Pearson Correlation Coefficient**.

Sinh viên tính ra sai số theo dạng **Mean Squared Error** để so sánh các giá trị tính ra và **các giá trị có thật** đối với 4 người dùng còn lại (***lưu ý chỉ so sánh với giá trị có thật trong dữ liệu***).

c) Câu 3 (2.0 điểm): Dự đoán giá chứng khoán.

Sinh viên sử dụng tệp **stockHVN2022.csv** cho câu này.

Bài toán đặt ra là cho giá chứng khoán 05 ngày liền trước của mã HVN, dự đoán giá trị của ngày hôm nay.

Sinh viên sử dụng dữ liệu từ tháng 01 đến hết tháng 08 để làm tập train, phần từ tháng 09 đến hết cho tập test.

Với mỗi lập sinh viên tạo ra một **DataFrame** có 2 cột

- **Giá 05 ngày trước**: một vector số thực chứa giá của 05 ngày trước
- **Giá hôm nay**: một số thực chứa giá của ngày hôm nay.

Ví dụ với chuỗi: a, b, c, d, e, f, g, h ta phát sinh được các mẫu dữ liệu (vế trái là giá 05 ngày trước, vế phải là giá hôm nay)

- *a, b, c, d, e* → *f*
- *b, c, d, e, f* → *g*
- *c, d, e, f, g* → *h*
- ...

Sinh viên lưu xuống hai **DataFrame** với tên dễ hiểu, hợp lý sau đó xây dựng mô hình **Linear Regression (pyspark)** để dự đoán giá chứng khoán theo bài toán trên.

Sử dụng tập train để huấn luyện và tập test để kiểm định. Sinh viên báo cáo kết quả sai số **Mean Squared Error** trên tập train và test.

d) Câu 4 (3.0 điểm): Phân loại đa lớp với pyspark

Sinh viên sử dụng kết quả của mục a) để làm câu này.

Với giá trị **r**, ta có thêm một tập dữ liệu train và test tương ứng. Có tất cả 02 bộ dữ liệu (train, test) gồm bộ dữ liệu gốc và bộ ứng với giá trị **r**.

Sinh viên xây dựng mô hình phân loại đa lớp với pyspark để nhận dạng ảnh thú cưng

- *Input: vector ảnh*
- *Output: chủng loại*
- *Hàm mục tiêu: Cross Entropy*
- *Độ đo: Accuracy.*

Sinh viên thực hiện huấn luyện và kiểm định mô hình đề ra với 02 bộ dữ liệu trên. Vẽ biểu đồ cột so sánh độ chính xác trên tập train và test cho 02 bộ dữ liệu (vẽ chung một biểu đồ) sử dụng thư viện **matplotlib.pyplot**.

e) Câu 5 (1.0 điểm): Báo cáo

- Sinh viên viết báo cáo, xuất thành tập tin ***report.pdf***. Trong đó, bao gồm các nội dung
    - Danh sách sinh viên: MSSV, Họ tên, Email, Phân công công việc, Mức độ hoàn thành.
    - Với mỗi yêu cầu, sinh viên trình bày hướng xử lý, cấu trúc mô hình, các biểu đồ, bảng thống kê kết quả và nhận xét.
    - **Độ dài tối đa cho báo cáo là 04 trang A4, font Times New Roman, fontSize 13, giãn dòng 1.5**

III. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp <MSSV1>_<MSSV2> (tương tự cho nhóm gồm 03 sinh viên), trong đó gồm:

- <MSSV1>_<MSSV2>.ipynb chứa mã nguồn đồ án
- <MSSV1>_<MSSV2>.pdf kết xuất pdf mã nguồn đồ án từ Google Colab.
- report.pdf: báo cáo đồ án.

- Lưu ý giữ lại kết quả thực thi của các ô trong cả hai tập tin .ipynb và .pdf.
- Nén thư mục thành <MSSV1>_<MSSV2>.zip và nộp theo deadline.

IV. Thời gian nộp bài

- Thời hạn nộp bài được thông báo trên hệ thống.
- Sinh viên nộp trễ hơn hạn trên đồng nghĩa với việc **0.0 điểm**.

V. Quy định

- **Mọi hành vi sao chép code trên mạng, chép bài bạn hoặc cho bạn chép bài nếu bị phát hiện đều sẽ bị điểm 0.0.**
- **Nếu bài làm của sinh viên có dấu hiệu sao chép trên mạng hoặc sao chép nhau, sinh viên sẽ được gọi lên phỏng vấn code để chứng minh bài làm là của mình.**

-- HẾT --