Ton Duc Thang University
Faculty of Information Technology

ĐỒ ÁN CUỐI KỲ

Môn: Xử lý dữ liệu lớn

Thời gian làm bài: 07 tuần

I. Hình thức

- Đề tài giữa kỳ được thực hiện theo nhóm **03** – **05** sinh viên.
- Nhóm sinh viên thực hiện các yêu cầu và nộp bài theo hướng dẫn 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**
cifar10-train-5k.csv	Dữ liệu hình ảnh và chủng loại tương ứng trong tập CIFAR-10[1]. 5000 dòng dữ liệu. Mỗi dòng chứa 1025 số nguyên • Số thứ nhất: loại hình ảnh (0, 1, 2, 3, …, 9) • 1024 số còn lại là pixel của ảnh grayscale 32 x 32.
cifar10-test-1k.csv	Tương tự như **cifar10-train-5k.csv** nhưng có 1000 dòng.
ratings2k.csv	Dữ liệu đánh giá sản phẩm Dòng 1 là header • **index**: chỉ số dòng • **user**: mã người dùng • **item**: mã món hàng • **rating**: đánh giá (0.0-5.0) 2365 dòng tiếp theo là dữ liệu tương ứng

[1] https://www.cs.toronto.edu/~kriz/cifar.html

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 219 dòng còn lại là dữ liệu tương ứng

a) Câu 1 (2.0 điểm): Phân cụm dữ liệu

Sinh viên sử dụng tập dữ liệu **cifar10-train-5k.csv** cho câu này.

Sử dụng **DataFrame** của **pyspark.sql** để khai thác dữ liệu và dùng thư viện **matplotlib.pyplot** để vẽ các biểu đồ trực quan.

Thí nghiệm với thuật toán k-Means (**pyspark.ml.clustering.KMeans**) và giá trị k trong đoạn [2; 10]:

- Tính trung bình khoảng cách từ mỗi điểm dữ liệu tới centroid tương ứng.
- Vẽ biểu đồ cột để so sánh các giá trị trung bình khoảng cách ở trên.
- Cho biết đâu là giá trị k cho trung bình khoảng cách thấp nhất.

Lưu ý: tổ chức mã nguồn theo mô hình hướng đối tượng.

b) Câu 2 (2.0 điểm): Giảm số chiều với SVD

Sinh viên sử dụng tập dữ liệu **cifar10-train-5k.csv** và **cifar10-test-1k.csv** cho câu này.

Sinh viên sử dụng thư viện **pyspark** và thuật toán **SVD** để giảm số chiều các vector ảnh trong mỗi tập xuống còn 64 (8 x 8).

Lưu kết quả thành **cifar10-train-svd.csv** và **cifar10-test-svd.csv**.

Lưu ý: tổ chức mã nguồn theo mô hình hướng đối tượng.

c) Câu 3 (2.0 điểm): Khuyến nghị sản phẩm với Collaborative Filtering

Sinh viên sử dụng tập **ratings2k.csv** cho câu này.

Sử dụng **pyspark** và thuật toán **Collaborative Filtering** để đọc dữ liệu.

Sinh viên cài đặt hàm

predictRating(df, userId, itemId, N)

để tìm ra giá trị rating mà người dùng *userId* dành cho sản phẩm *itemId*.

- Nếu đã có giá trị rating [*userId, itemId*] thì trả về giá trị đó.
- Nếu chưa có giá trị rating cần tìm, dùng thuật toán **Collaborative Filtering** để dự đoán.
- Nếu giá trị *userId* hoặc *itemId* không hợp lệ thì trả về -1.

Các tham số trong hàm:

- *df*: **DataFrame** dữ liệu được cho
- *userId*: mã người dùng cần tìm
- *itemId*: mã món hàng cần xử lý
- *N*: số lượng người "giống" với *userId* nhất đã đánh giá món hàng *itemId* (dùng trong thuật toán **Collaborative Filtering**).

Lưu ý: tổ chức mã nguồn theo mô hình hướng đối tượng.

d) Câu 4 (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 *k* ngày liền trước của mã HVN, dự đoán giá trị của ngày tiếp theo.

Sinh viên sử dụng dữ liệu từ tháng 01 đến hết tháng 06 để làm tập train, phần từ tháng 07 đến hết cho tập test.

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

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

Với giá trị k trong đoạn [5; 10], sinh viên

- Xây dựng mô hình **Linear Regression (pyspark)** để dự đoán giá chứng khoán theo bài toán trên: học dữ liệu từ tập training và đánh giá trên tập test.

- Tính ra sai số **Mean Square Error** trên tập training và test với mô hình đã huấn luyện.
- Sử dụng **matplotlib.pyplot** vẽ biểu đồ cột thể hiện giá trị **Mean Square Error** trên tập training và test tìm được ứng với các giá trị *k*.

Lưu ý: tổ chức mã nguồn theo mô hình hướng đối tượng.

e) Câu 5 (1.0 điểm): Phân loại đa lớp với pyspark

Sử dụng tập các tập dữ liệu sau:

- **cifar10-train-5k.csv** (*train*)
- **cifar10-test-1k.csv** (*test*)
- **cifar10-train-svd.csv** (*train-svd*)
- **cifar10-test-svd.csv** (*test-svd*)

Tổng cộng có hai bộ dữ liệu gồm (*train*, *test*) ban đầu và (*train-svd*, *test-svd*) từ câu b). Sinh viên xây dựng mô hình phân loại đa lớp với **pyspark**

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

Sinh viên tìm hiểu và áp dụng ba mô hình phân lớp thông dụng trong pyspark gồm:

- Multi-layer Perceptron https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier
- Random Forest https://spark.apache.org/docs/latest/ml-classification-regression.html#random-forest-classifier
- Linear Support Vector Machine: https://spark.apache.org/docs/latest/ml-classification-regression.html#linear-support-vector-machine

Sinh viên vẽ **biểu đồ cột tứ** với matplotlib.pyplot để thể hiện độ chính xác của ba mô hình trên bốn tập dữ liệu.

Lưu ý: tổ chức mã nguồn theo mô hình hướng đối tượng.

f) Câu 6 (1.0 điểm): Báo cáo

- Sinh viên viết báo cáo kết quả thực hiện đề tài.
- **KHÔNG CÓ MẪU BÁO CÁO, NHÓM SINH VIÊN TỰ TỔ CHỨC NỘI DUNG**.
- Các thông tin tối thiểu cần có:
  - Danh sách sinh viên: MSSV, Họ tên, Email, Phân công công việc, Mức độ hoàn thành.
  - Tóm tắt cách xử lý từng yêu cầu, nên diễn đạt bằng mã giả/sơ đồ.
  - HẠN CHẾ TỐI ĐA NHÚNG MÃ NGUỒN THÔ VÀO BÀI THUYẾT TRÌNH.
  - Các nội dụng tìm hiểu cần trình bày cô đọng, có ví dụ trực quan.
  - Thuận lợi và khó khăn trong đề tài.
  - Bảng tự đánh giá mức độ hoàn thành các yêu cầu.
  - Tài liệu trích dẫn ghi theo định dạng IEEE.
- Yêu cầu về định dạng: hạn chế dùng nền tối, đảm bảo khi in dạng trắng đen thì các nội dung vẫn rõ ràng.

III. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp

  CK_<Mã nhóm>

  trong đó gồm:
  - **source.ipynb** → chứa mã nguồn đồ án (giữ lại các kết quả chạy)
  - **source.pdf** → kết xuất pdf của notebook
  - **report.pdf** → bài thuyết trình.
- Nén thư mục thành tệp zip và nộp theo deadline.

IV. Quy định

- **Nhóm sinh viên nộp trễ hạn bị 0.0 điểm toàn nhóm.**
- **Thiếu sót các tài liệu được yêu cầu trong tệp nộp bài sẽ bị trừ tối thiểu 50% điểm phần báo cáo.**
- **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 riêng để chứng minh bài làm là của mình.**

-- HẾT –