Ton Duc Thang University
Faculty of Information Technology

ĐỒ ÁN CUỐI KỲ

Môn: Xử lý dữ liệu lớn

Thời gian làm bài: 05 tuần

I. Hình thức

- Đề tài giữa kỳ được thực hiện theo nhóm **04** – **05** sinh viên.
- 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**
mnist_small_train.csv	Dữ liệu hình ảnh và chủng loại chữ số viết tay MNIST[1]. 7000 dòng dữ liệu. Mỗi dòng chứa 785 số nguyên • Số thứ nhất: loại chữ số (0, 1, 2, 3, …, 9) • 784 số còn lại là pixel của ảnh grayscale 28 x 28.
mnist_small_test.csv	Tương tự như **mnist_small_train.csv** nhưng có 3000 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://en.wikipedia.org/wiki/MNIST_database

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 **mnist_small_test.csv** cho câu này.

YC1_1: Hiển thị

Sử dụng **DataFrame** của **pyspark.sql** và thư viện **matplotlib.pyplot** để vẽ ra biểu đồ 3 x 5 ô hiển thị 15 bức ảnh đầu tiên.

- Chuyển đổi vector ảnh 784 chiều thành ma trận 28 x 28
- Hiển thị ảnh với hàm **imshow**() của matplotlib.pyplot
- Với từng bức ảnh trong lưới 3 x 5, hiển thị **title** là label (chủng loại) của chữ số.

YC1_2: Phân cụm

Sử dụng thư viện **pyspark.ml.clustering.KMeans** để tiến hành phân cụm các vector ảnh lần lượt với 3 giá trị k là 5, 10, 15.

Với mỗi thí nghiệm (một giá trị k) tiến hành:

- Tạo ra mô hình
- Phân cụm với k và khoảng cách Euclidean.
- Lưu mô hình xuống file
- Load mô hình từ file
- Tính ra tổng khoảng cách Euclidean từ mỗi điểm dữ liệu tới centroid tương ứng.

YC1_3: Trực quan hoá kết quả

Sử dụng thư viện matplotlib.pyplot để vẽ biểu đồ cột thể hiện giá trị tổng tổng khoảng cách Euclidean từ mỗi điểm dữ liệu tới centroid tương ứng cho các giá trị k ở trên.

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

YC2_1: Giảm số chiều tập train

Sinh viên sử dụng thư viện **pyspark** và thuật toán **SVD** để giảm số chiều các vector ảnh trong tập **mnist_small_train.csv** xuống còn 196 (14 x 14). Sau đó lưu lại kết quả thành 01 tệp csv có cấu trúc tương tự **mnist_small_train_svd.csv**.

YC2_2: Giảm số chiều tập test

Thực hiện tương tự **YC2_1** trên tập mnist_small_test.csv và lưu kết quả xuống tập tin **mnist_small_test_svd.csv**.

c) Câu 3 (1.0 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 60 người dùng đầu tiên, bằng phương pháp **Collaborative Filtering**, tính ra tất cả ratings cho 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 người dùng còn lại (***lưu ý chỉ so sánh với giá trị có thật trong dữ liệu***).

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 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 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 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: học dữ liệu từ tập training và đánh giá trên tập test.
- Lưu mô hình xuống tập tin
- Đọc mô hình từ tập tin
- 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.

e) Câu 5 (2.0 điểm): Phân loại đa lớp với pyspark

Sử dụng tập **mnist_small_training/test.csv** và **mnist_small_training/test_svd.csv** cho câu này.

Tổng cộng có hai bộ dữ liệu gồm (training, test) ban đầu và (training_svd, test_svd) từ câu b).

Sinh viên xây dựng mô hình phân loại đa lớp với **pyspark** để nhận dạng ảnh chữ số

- *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 MNIST gồm training, test, training_svd và test_svd.

4 tập dữ liệu x 3 mô hình

f) Câu 6 (1.0 điểm): Báo cáo

- Sinh viên viết báo cáo kết quả đề tài theo hình thức thuyết trình. **KHÔNG CÓ MẪU THUYẾT TRÌNH, 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: tỷ lệ slide 4x3, hạn chế dùng nền tối/màu sắc vì máy chiếu mờ, đảm bảo khi in bài thuyết trình dạng trắng đen thì các nội dung vẫn rõ ràng.
- Thời lượng tối đa cho phần thuyết trình là **05 phút**.
- ***Hướng dẫn quay video đính kèm trong phụ lục.***

III. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp
  <MSSV1>_<MSSV2>_<MSSV3>_<MSSV4>_<MSSV5>
  trong đó gồm:
  - **source.ipynb** → chứa mã nguồn đồ án (giữ lại các kết quả chạy)
  - **source.pdf** → kết xuất pdf của notebook
  - **presentation.pdf** → bài thuyết trình.
  - **video.txt** → URL tới video 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.**
- **Sai sót mã số sinh viên nào trong tên tệp nộp bài thì sinh viên tương ứng bị 0.0 điểm.**
- **Thiếu sót các tài liệu được yêu cầu trong tệp nộp bài sẽ bị trừ tối thiểu 50% điểm phần thuyết trình.**
- **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 –

HƯỚNG DẪN
VIDEO THUYẾT TRÌNH
INSTRUCTIONS
FOR RECORDING PRESENTATION VIDEO

I. Mục tiêu/*Objectives*

- Nhóm sinh viên thực hiện quay video thuyết trình để báo cáo nội dung đồ án/đề tài.

  Student groups record a video to present your project/topic.
- Hình thức, công cụ, thời lượng được mô tả chi tiết ở mục tiếp theo.

  Formation, tools, and duration are described in the next section.

II. Yêu cầu/*Requirements*

- Công cụ: Zoom/Google Meet

  Tools: Zoom/Google Meet
- Thời lượng: theo mô tả trong đồ án

  Duration: designated in the project
- Hình thức:

  Formation
  - Nhóm sinh viên tạo một meeting để tham gia thuyết trình.

    Student groups create a meeting to present your work.
  - Đặt tên hiển thị theo dạng <MSSV>_<Họ tên>,

    Set the display name as <Student ID>_<Full Name>

    ví dụ 52200001_Nguyễn Văn A

    for example, 52200001_Nguyen Van A
  - Tất cả thành viên phải bật camera trong toàn bộ buổi thuyết trình.

    Every member must turn on your camera during the presentation.

- Sinh viên trình chiếu bài thuyết trình nhưng phải đảm bảo hiển thị đầy đủ khuôn mặt của các thành viên còn lại.
  Students show your presentation but ensuring to display all member facial thumbnails.
- Các sinh viên thay phiên trình bày các nội dung.
  Students share the role of presenting.
- Bài thuyết trình được quay lại, đảm bảo chất lượng hình ảnh và âm thanh.
  The presentation is recorded completely.

III. Hướng dẫn nộp bài/Submission Instructions

- Video tải lên Youtube và đặt ở chế độ "unlisted", tuyệt đối không để dạng "public".
  The recording is uploaded to Youtube with the "unlisted" sharing option. Do not share with the "public" option.
- Đặt tên theo cú pháp
  Rename the recording as below

  <Năm học>-<Học kỳ>-<Môn học>-<Tên nhóm>

  <Year>-<Semester>-<Course>-<Group Name>

  trong đó gồm:
  - <Năm học> theo dạng YYYY, ví dụ 2223, 2324, 2425
    <Year> in form of YYYY, for example, 2223, 2324, 2425
  - <Học kỳ> là "HK1" hoặc "HK2"
    <Semester> is "HK1" (Term 1) or "HK2" (Term 2)
  - <Môn học> là "AI" (NM Trí tuệ Nhân tạo) hoặc "MMDS" (Xử lý Dữ liệu lớn)
    <Course> is "AI" (Introduction to AI) or "MMDS" (Mining Massive Datasets)
  - <Tên nhóm> theo tên đã đăng ký
    <Group Name> as in registration
- Một thành viên nhóm đại diện nộp đường dẫn theo deadline được cho.

Only one representative student submits the video URL by the deadline.

-- HẾT --

-- END --