Ton Duc Thang University
Faculty of Information Technology

# ĐỒ ÁN CUỐI KỲ

## **Môn**: Xử lý dữ liệu lớn

## **Thời gian làm bài**: 06 tuần

## I. Hình thức

- Đồ án được thực hiện theo nhóm **04** – **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** |
|---|---|
| **mnist_mini.csv** | Dữ liệu hình ảnh ký số viết tay trong tập MNIST.<br>10000 dòng dữ liệu.<br>Mỗi dòng chứa 785 số nguyên<br>• Số thứ nhất: loại ký số (0, 1, 2, 3, …, 9)<br>• 784 số còn lại là pixel của ảnh grayscale 28 x 28. |
| **ratings2k.csv** | Dữ liệu đánh giá sản phẩm<br>Dòng 1 là header<br>• **index**: chỉ số dòng<br>• **user**: mã người dùng<br>• **item**: mã món hàng<br>• **rating**: đánh giá (0.0-5.0)<br>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).<br>Dòng 1 là header:<br>• **Ngay**: ngày ghi nhận<br>• **HVN**: giá đóng cửa<br>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_mini.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.

Cài đặt thuật toán k-Means (**pyspark.ml.clustering.KMeans**) và giá trị **k = 10**, trong đó các điểm dữ liệu tại dòng **0, 1, 2, 3, 4, 7, 8, 11, 18, 61** được gán trọng số gấp **100** lần các điểm dữ liệu khác.

Với mỗi cluster, tính trung bình khoảng cách từ các điểm dữ liệu tới centroid. Vẽ biểu đồ cột để trực quan hoá kết quả.

*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 **mnist_mini.csv** cho câu này.

Sử dụng thư viện **pyspark** và thuật toán **SVD** để giảm số chiều các điểm dữ liệu từ 784 xuống 3.

Chọn ngẫu nhiên **100** điểm dữ liệu sau khi giảm số chiều. Sử dụng kết quả phân cụm ở câu 1 để vẽ biểu đồ 3D mô tả phân bố của **100** điểm này trong không gian với thư viện **matplotlib.pyplot**.

*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.

Chia tập dữ liệu thành tập **training** và tập **test** với tỷ lệ **7 : 3.**

Sử dụng **pyspark** và thuật toán **ALS** để khảo sát hiệu suất của mô hình theo độ đo **Mean Squared Error (MSE)** với các giá trị số lượng người dùng "tương đồng" trong đoạn **[10; 20].**

Chạy inference để minh hoạ hoạt động của mô hình.

Vẽ biểu đồ cột để trực quan hoá các giá trị sai số MSE.

*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 biên độ dao động giá chứng khoán *k* ngày liền trước của mã HVN, dự đoán biên độ 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 tập dữ liệu được cho, sinh viên tạo ra cột "**fluctuation**" chứa biên độ dao động của giá cổ phiếu theo công thức sau:

Biên độ ngày [k] = (Giá ngày [k] – Giá ngày [k-1]) / Giá ngày [k-1]

Ngày đầu tiên trong dữ liệu có biên độ dao động là **0.0%.**

Sau đó, sinh viên phát sinh ra một **DataFrame** có 2 cột

- **Biên độ 5 ngày trước**: một vector số thực chứa biên độ của 05 ngày trước
- **Biên độ ngày tiếp theo**: một số thực chứa biên độ của ngày hôm nay.

Xây dựng mô hình **Linear Regression (pyspark)** để dự đoán biên độ 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.

*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 tập dữ liệu **mnist_mini.csv** sau:

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 đôi với matplotlib.pyplot** để thể hiện độ chính xác của ba mô hình trên tập training và test.

*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** → báo cáo.

- 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.**
- **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 –**