Chủ đề cho Assignment HK212 (3 tuần)

Chủ đề 1 (3 weeks): Streaming/Big data clustering (tập trung vào large/very large dataset clustering)

- Kỹ thuật: improved versions/variants k-means, BIRCH, CLARAN, etc. (key idea, mô tả kỹ thuật, đưa ra Pros/Cons, tính phổ dụng).
- Hiện thực (implementation), hoặc học và sử dụng một số well-know tools, frameworks.
- Một số ứng dụng trong việc phát hiện bất thường.
- Thực nghiệm (benchmark datasets, artifficially generating datasets, e.g. HAWKS generator).
- Phân tích, đánh giá (Adjust Rand Index, Budin index, hoặc các các index khác).

Chủ đề 2 (3 weeks): Highly imbalanced classification using sampling

- Model: Neural Nets, CNN và các biến thể liên quan đến deep learning (key idea, mô tả kỹ thuật, đưa ra Pros/Cons, tính phổ dụng).
- Tìm hiểu một số kỹ thuật về sampling (upper/under sampling, ví dụ như SMOTE).
- Một số ứng dụng trong việc phát hiện bất thường.
- Hiện thực.
- Thực nghiệm (Datasets: https://www.kaggle.com/mlg-ulb/creditcardfraud, https://arxiv.org/pdf/1106.1813.pdf, …)
- Phân tích và đánh gia (accuracy, confusion matrix)

Nhóm: 3-4 người, tự chia, gửi lại danh sách cho **NGUYỄN HOÀI THƯƠNG** (thuong.nguyenk_19@hcmut.edu.vn) chậm nhất là 12/04/2022 danh sách gồm Nhóm và Chủ đề tương ứng.

Báo cáo:

- Thời gian bắt đầu BTL là: 11/04/2022.
- Report (Giới thiệu, mô tả, thực nghiệm và kết quả, phân), Presentation (ngắn gọn: mô tả, kết quả thực nghiệm, phân tích).
- Gửi qua email với tiêu đề [KPDL-202-BTL] – Nhóm <?> đến lhtrang@hcmut.edu.vn