KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
ĐỒ ÁN THỰC HÀNH

I. Quy định

- Đồ án làm theo nhóm, mỗi nhóm có tối đa 2 sinh viên.
- Hạn đăng ký: xem trên trang web môn học.
- Các nội dung cần nộp bao gồm:
  - ***Báo cáo***: tập tin ***.pdf** và ***.docx** (*Chỉ nộp tập tin, không in*).
  - ***Thuyết trình***: tập tin ***.pptx** (*Chuẩn bị slides thuyết trình trong 15 phút*).
  - ***Chương trình***: source code của chương trình (*Source code phải biên dịch thành công*).
  - ***Hướng dẫn sử dụng***: tập tin ***.txt** (*Ghi rõ thông tin nhóm, các thư viện cần có, cách biên dịch và các chức năng hỗ trợ*).

II. Nội dung yêu cầu

1. Báo cáo

Nội dung báo cáo tối đa 8 trang A4 (*không tính phần phụ lục, định dạng font Times New Roman và size 13*) bao gồm:

- Thông tin của nhóm, môn học, chủ đề đồ án môn học.
- Mục lục.
- Mô tả tóm tắt về tập dữ liệu.
- Các bước tiền xử lý dữ liệu (*nếu có*).
- Các thuộc tính được sử dụng: số lượng thuộc tính và ý nghĩa từng thuộc tính.
- Liệt kê phương pháp sử dụng trong chương trình gồm: tên phương pháp, các đặc trưng chính của phương pháp đó một cách ngắn gọn, thuật toán mỗi phương pháp (*tối đa 2 trang A4*).
- Liệt kê các thực nghiệm đã tiến hành theo định dạng sau:

- **Thực nghiệm 1:**
  - Dữ liệu huấn luyện: số lượng mẫu.
  - Dữ liệu kiểm tra 1: số lượng mẫu.
  - Dữ liệu kiểm tra 2: số lượng mẫu.

- …
- Kết quả thực nghiệm:

Kiểm tra	Precision	Recall	F-measure
1	0.9	0.87	…
2	…	…	…

- Biểu đồ theo F-measure.
- Nhận xét kết quả thực nghiệm và biểu đồ.

- **Thực nghiệm 2:**
  - …
  -

2. *Thuyết trình*

- Chuẩn bị slides thuyết trình trong 15 phút bao gồm các nội dung: *thông tin nhóm, bố cục trình bày, giới thiệu, nội dung trình bày, kết luận, demo*.
- Tất cả thành viên trong nhóm đều tham gia thuyết trình.
- Tiến hành chạy thực nghiệm trực tiếp trên lớp.

3. *Chương trình*

Chương trình cần có các chức năng cơ bản sau:

- Huấn luyện mô hình phân lớp.
- Phân lớp cho một mẫu bất kỳ (*nhập từ bàn phím*).
- Phân lớp cho nhiều mẫu (*nhập thông qua tập tin*).
- Đánh giá kết quả phân lớp theo độ đo Precision, Recall và F-measure.

Các điều cần lưu ý:

- Định dạng của tập tin đầu vào do sinh viên quyết định.
- Source code cần được viết rõ ràng có chú thích (*khuyến khích viết theo hướng đối tượng*).
- Sinh viên được phép sử dụng thư viện có sẵn, hỗ trợ cài đặt các thuật toán phân lớp, không bắt buộc phải tự cài đặt. Một đề xuất cho sinh viên là bộ thư viện cung cấp bởi WEKA, viết trên nền ngôn ngữ JAVA.
  - http://weka.wikispaces.com/
  - http://weka.wikispaces.com/Use+WEKA+in+your+Java+code

III. Danh sách các đồ án

1. Tiếp thị ngân hàng

- Link dữ liệu: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- Mục tiêu: Xây dựng mô hình dự đoán một khách hàng có tham gia gửi tiền có kỳ hạn vào ngân hàng hay không.

2. Phân lớp dáng bộ người dùng

- Link dữ liệu:
  - http://archive.ics.uci.edu/ml/datasets/Wearable+Computing%3A+Classification+of+Body+Postures+and+Movements+%28PUC-Rio%29
  - groupware.les.inf.puc-rio.br/static/har/dataset-har-PUC-Rio-ugulino.zip
- Tài liệu tham khảo:
  - http://groupware.les.inf.puc-rio.br/har#ixzz2PyRdbAfA
- Mục tiêu: Xây dựng mô hình phân lớp dáng bộ của một người dùng thông qua các thông tin liên quan đến hoạt động của họ. Bao gồm 5 lớp chính như sau: *ngồi xuống, đứng lên, đứng yên, đi bộ và đang ngồi*.

3. Poker Hand

- Link dữ liệu: http://archive.ics.uci.edu/ml/datasets/Poker+Hand
- Mô tả: Bài poker sử dụng các thuật ngữ như bộ bài Tây 52 lá. Mỗi người chơi sẽ được chia 5 lá, người thắng là người có bài cao nhất. Dưới đây là bảng xếp hạng bài trong poker (*tăng dần*):
  - Mậu: không có gì cả.
  - Dách: có 1 đôi.
  - Thú: có 2 đôi.
  - Xám chi: có 3 lá bài cùng hạng.
  - Sảnh: có 5 lá bài liên tiếp.
  - Thùng: có 5 lá bài cùng chất.
  - Cù lũ: có 3 lá bài cùng hạng và 1 đôi.
  - Tứ quý: có 4 lá bài cùng chất.

- Thùng phá sảnh: có 5 lá bài liên tiếp và cùng chất.
- Mậu binh: là thùng phá sảnh với 5 lá (10, J, Q, K, A).

– Mục tiêu: Xây dựng mô hình để nhận biết 5 lá bài được chia cho một người thuộc hạng nào trong 10 hạng trên.

4. Phân loại tin nhắn rác

– Link dữ liệu:
  - http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

– Mục tiêu: Xây dựng bộ lọc có thể nhận biết được đâu là tin nhắn rác (*spam*) và không phải tin nhắn rác (*ham*) để tự động loại bỏ các tin nhắn rác trong hợp thư chính.

5. Quảng cáo nông trại

– Link dữ liệu: http://archive.ics.uci.edu/ml/datasets/Farm+Ads

– Mục tiêu: Dữ liệu là các mẫu quảng cáo liên quan đến nông trại. Các mẫu quảng cáo này được đăng trên các trang web khác nhau bởi Ad Networks. Tuy nhiên, có các mẫu quảng cáo không phù hợp với người dùng của trang. Hãy xây dựng mô hình để giúp người dùng phân loại tự động các mẫu quảng cáo nào phù hợp (+1) và không phù hợp (–1).