LAB03

MÁY HỌC

1 Yêu cầu

- Nội dung: Sinh viên tìm hiểu công cụ Weka và trải nghiệm các chức năng để chạy các thuật toán trong phần Máy Học.
- Dạng bài tập: nhóm 3 người.
- Thời gian: 2 tuần
- Nộp bài: tất cả nội dung được nén lại và nộp trên moodle.

2 Nội dung chi tiết

2.1 Tìm hiểu công cụ Weka (30%)

- Tìm hiểu công cụ Weka gồm giải thích các chức năng, cách sử dụng ở mức cơ bản. Viết báo cáo ở dạng Word. Tối thiểu 10 trang. Khuyến khích sử dụng hình ảnh, ví dụ minh họa.

2.2 Sử dụng Weka để chạy thuật toán ID3 (30%)

Cho tập dữ liệu: Zoo (http://archive.ics.uci.edu/ml/datasets/Zoo) - tập dữ liệu về động vật.

Thực hiện:

- Tạo tập tin Zoo.arff chứa dữ liệu Zoo.
- Hãy mô tả tổng quát về dữ liệu Zoo:
  - Số mẫu
  - Tên và ý nghĩa các thuộc tính
  - Danh sách các phân lớp. Hãy đặt tên ngắn gọn cho mỗi phân lớp và chỉnh sửa file Zoo.arff sao cho thuộc tính phân lớp gồm các tên mới này thay vì các con số từ 1 đến 7 như trong dữ liệu thô.
- Sử dụng thuật toán ID3 để học ra cây quyết định từ dữ liệu trên (cách phân chia dữ liệu học là tùy ý).
- Báo cáo cây đã sinh ra bởi quá trình chạy.
- Với cây đã sinh ra ở trên, cho biết kết quả cho 5 mẫu sau đây:

- 1. NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1, ?
- 2. NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0, ?
- 3. NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0, ?
- 4. NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0, ?
- 5. NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0, ?

2.3 Chạy các thuật toán khác (40%)

1. (10%) Hãy viết chương trình bằng python cho giải thuật Naïve Bayes rồi chạy tương tự cho các yêu cầu của bài 2.2 ở trên. (lưu ý phải báo cáo hàm làm trơn được áp dụng cho code của các bạn là gì?)
2. Bây giờ quay trở lại dùng weka, SV chạy thêm các thuật toán đã học trên lớp hoặc tìm hiểu thêm như Naïve Bayes, … và báo cáo các kết quả xuất ra. Trình bày quá trình thiết lập và chạy dữ liệu kèm theo hình ảnh minh họa. (Mỗi thuật toán 10%).

3 Qui định

- Hạn nộp: ***xem trên Moodle.***
- Đặt tên chương trình là MSSV1_MSSV2_Lab03, với MSSV là mã số sinh viên. ***Report***: chứa tập tin báo cáo (.doc, .docx hoặc pdf) trình bày các kiến thức đã được yêu cầu, các minh chứng về chạy dữ liệu.

**** Lưu ý: Các bài làm giống nhau sẽ bị 0 điểm.***