

BÀI TẬP THỰC HÀNH 4

CHỦ ĐỀ: HỌC MÁY

1 Yêu cầu

1.1 Tìm hiểu công cụ Weka

- 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 20 trang. Khuyến khích sử dụng hình ảnh, ví dụ minh họa.

1.2 Sử dụng Weka để chạy thuật toán ID3

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:
 - o Số mẫu
 - o Tên và ý nghĩa các thuộc tính
 - o 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:
 - o 1. NameIsSecret,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1, ?
 - o 2. NameIsSecret,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0, ?
 - o 3. NameIsSecret,0,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0, ?
 - o 4. NameIsSecret,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0, ?
 - o 5. NameIsSecret,0,0,1,0,0,1,1,1,1,1,0,0,4,1,0,0, ?

2 Mở rộng

SV sẽ được điểm cộng nếu làm thêm các phần mở rộng sau đây:

- a) Báo cáo về Weka chi tiết với hàm lượng tri thức cao
- b) Chạy thuật toán Bayes và báo cáo các kết quả xuất ra

3 Qui định

- Dạng bài tập: *cá nhân*
- Hạn nộp: *xem trên Moodle.*
- Đặt tên chương trình là MSSV1_Lab04, 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.**