LAB03

MÁY HỌC

1 Yêu cầu

- Nội dung:
  - Nội dung 1 (tỉ lệ 50%): Sinh viên tìm hiểu công cụ Weka và trải nghiệm các chức năng để chạy các thuật toán trong phần Máy Học.
  - Nội dung 2 (tỉ lệ 50%): Sinh viên sử dụng ngôn ngữ C++/C#/Java/Python để cài đặt lại thuật toán máy học (không sử dụng thư viện).
- Dạng bài tập: nhóm 2 người.
- Thời gian: 3 tuần
- Nộp bài: tất cả nội dung được nén lại và nộp trên moodle.

2 Nội dung 1 (50%)

2.1 Tìm hiểu công cụ Weka (40%)

- 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 (60%)

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, ?

3 Nội dung 2 (50%)

3.1 Tải tập dữ liệu

SV tải về một số tập dữ liệu trong csdl sau:

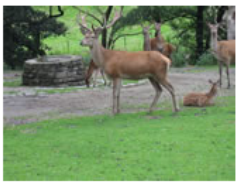
http://archive.ics.uci.edu/ml/datasets.html

SV nên chọn các dữ liệu có đặc điểm là dữ liệu rời rạc/số, bài toán phân lớp, không có dữ liệu thiếu. Ví dụ như tập dữ liệu Zoo.

Zoo Data Set

Download: Data Folder, Data Set Description

Abstract: Artificial, 7 classes of animals



Data Set Characteristics:	Multivariate	**Number of Instances:**	101	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	17	**Date Donated**	1990-05-15
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	187157

3.2 Cài đặt thuật toán ID3

Viết chương trình thực hiện các công việc sau:

- Đọc dữ liệu và thuộc tính từ một trong các tập dữ liệu trên (không bắt buộc làm hết mọi tập dữ liệu).

- Chia tập dữ liệu ra làm 2 phần: phần học và phần kiểm thử.
- Xây dựng cây quyết định dựa trên tập dữ liệu và xuất ra màn hình console tương tự như trong Weka.
- Đánh giá độ chính xác của thuật toán bằng cách sử dụng cây quyết định để tìm các nhãn trên tập thử.

Sinh viên có thể tham khảo code đọc và xây cây Id3 trên mạng nhưng cần phải tự code lại và điều chỉnh để phù hợp với dữ liệu, ghi rõ nguồn tài liệu tham khảo.

4 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.***