

Bài tập Học máy

Câu 1

Cho bảng quan sát sau:

STT	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
1	Nhỏ	Đức	Độc thân	A
2	Lớn	Pháp	Độc thân	A
3	Lớn	Đức	Độc thân	A
4	Nhỏ	Ý	Độc thân	B
5	Lớn	Đức	Có gia đình	B
6	Lớn	Ý	Độc thân	B
7	Lớn	Ý	Có gia đình	B
8	Nhỏ	Đức	Có gia đình	B
9	Nhỏ	Pháp	Có gia đình	?

Hãy sử dụng phương pháp cây định danh để xác định điều kiện cho Nhóm.

Câu 2

Cho bảng quan sát rút ra từ thị trường chứng khoán như sau:

Mẫu	Thời gian	Cạnh tranh	Loại	Lợi nhuận
1	Cũ	Có	Phần mềm	Giảm
2	Mới	Có	Phần mềm	Tăng
3	Trung bình	Không	Phần mềm	Tăng
4	Trung bình	Có	Phần mềm	Giảm
5	Mới	Không	Phần cứng	Tăng
6	Cũ	Không	Phần mềm	Giảm
7	Cũ	Không	Phần cứng	Giảm
8	Trung bình	Không	Phần cứng	Tăng
9	Trung bình	Có	Phần cứng	Giảm
10	Mới	Không	Phần mềm	Tăng
11	Mới	Có	Phần cứng	?

Hãy sử dụng thuật toán cây định danh để xác định điều kiện của việc Tăng hay Giảm của Lợi nhuận.

Câu 3

Từ các cây quyết định đã xây dựng, rút ra tập luật phân lớp và dự đoán cho các mẫu chưa có quyết định.

Câu 4

Một phiên bản khác của thuật toán ID3 sử dụng Informatic Gain thay cho entropy để chọn thuộc tính quyết định. Công thức tính Informatic Gain như sau:

$$\text{Gain}(A) = \text{Entropy}(S) - \text{Entropy}(A)$$

Trong đó:

S: tập mẫu
A: là một thuộc tính
Entropy(S): độ hỗn loạn của tập S
Entropy(A): độ hỗn loạn trung bình của thuộc tính A (theo công thức trên lớp)
Em có nhận xét gì về độ đo này (nó có tương tự với entropy hay không, lựa chọn thuộc tính tốt nhất như thế nào)? Vận dụng độ đo này để thực hiện các câu 1, 2.

Câu 5

Sử dụng thuật toán ILA và Naïve Bayes để thực hiện lại các bài tập 1, 2.