Bài tập

Câu I. Cho tập mẫu gồm các đối tượng có hai thuộc tính đặc trưng, đặc trưng thứ nhất nhận giá trị T / F, đặc trưng thứ hai nhận giá trị trong 5 chữ số {1,2,3,4,5}. Dữ liệu quan sát được gồm hai lớp cho bởi bảng sau:

ω_1 : (T,1) (T,2) (F,2) (T,4) (F,5)
ω_2 : (F,1) (T,3) (F,3) (F,4) (T,5)
Hãy xây dựng cây quyết định tương ứng nhờ thuật toán C4.5.

Câu II. Cho tập dữ liệu quan sát trong R^2 gồm 3 lớp:

$S_1 = \{((1.0), (0,1)\}$; $S_2 = \{,(1,2) (2,1)\}$; $S_3 = \{(-1,0), (-1,-1), (-2,-2), (0,-1)\}$

1)Áp dụng thuật toán học perceptron để phân biệt S_1 và S_2 với khởi tạo là:

$\mathbf{w} = (1,1)^T$ và $w_0 = 0$

2) Chỉ ra các hàm phân biệt cho mỗi lớp khi dùng phương pháp khoảng cách cực tiểu theo khoảng cách Mahalanobis có matrận A = $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ và khoảng cách Euclide

3) Chỉ xét hai lớp S_2 và S_3, hãy áp dụng quy tắc k-NN để phân lớp điểm (0, 0,4) khi dùng k=3 với khoảng cách xác định bởi chuẩn Trêbusep:

$$\|x-y\|_t = \max\left(|x_j - y_j|\right)$$

Câu III. Một loại máy của công ty có 0,8% máy phải bảo trì sự cố trong 1 năm. Điều tra cho thấy 97% máy bảo trì có sử dụng 1 loại linh kiện nhập từ Trung Quốc còn những máy tốt có 1% lắp loại linh kiện này. Một khách hàng định mua một máy có lắp loại linh kiện này

1) Dùng tiêu chuẩn MAP để quyết định xem có nên thay linh kiện này trước khi giao cho khách để tránh bảo trì không?
2) Nếu giá linh kiện này là 2 USD còn phí bảo là 20 USD thì có nên thay không?

Câu IV. Một địa phương có 210000 dân, phát hiện 30 người bị bệnh lạ. Trong số những người nhiễm bệnh, có 29 người bị sốt còn những người không nhiễm bệnh có 1% bị sốt. Một ngườimới bị sốt .

3) Dùng tiêu chuẩn MAP để đoán xem người này có nhiễm bệnh không?
4) Một liều thuốc điều trị sớm là 2 triệu đồng còn điều trị muộn là 200 triệu đồng, anh (chị) hãy quyết định xem có nên cho người này dùng thuốc không?

Câu 6. 1) Mô tả thuật toán CART xây dựng cây hồi quy và phân lớp cho trường hợp giá trị thuộc tính thực

2) Xây dựng cây quyết định nhờ dùng CART nếu dữ liệu quan sát của hai lớp có hai đặc trưng thu được như sau:

Mẫu	a_1	a_2	Lớp
x_1	*0,15*	*0,84*	ω_1

x_2	*0,8*	*0,54*	ω_1
x_3	*0,28*	*0,36*	ω_1
x_4	*0,38*	*0,70*	ω_1
x_5	*0,52*	*0,48*	ω_1
x_6	*0,58*	*0,74*	ω_1
x_7	*0,74*	*0,76*	ω_1
x_8	*0,08*	*0,14*	ω_2
x_9	*0,24*	*0,16*	ω_2
x_{10}	*0,70*	*0,20*	ω_2
x_{11}	*0,90*	*0,28*	ω_2
x_{12}	*0,64*	*0,90*	ω_2
x_{13}	0,74	0,36	ω_2
x_{14}	0,10	0,30	ω_2

3)Chứng minh rằng kết quả của các thuật toán Find-S và loại trừ ứng cử không phụ thuộc vào thứ tự lấy các mẫu quan sát.