Đề thi giữa học kỳ (4/2022)
Môn : Học Máy và Ứng dụng (Cao học)
(Hạn chót nộp bài làm: 12:00 giờ PM ngày 2/4/2022
Địa chỉ email nộp bài: dtanhcse@gmail.com
Bài làm phải là: 1 file word hoặc 1 file PDF)

Đề thi gồm 2 trang

1. (0.75 điểm)
Cho biết loại dữ liệu (liên tục hay rời rạc) phù hợp với từng giải thuật phân lớp sau đây:
- giải thuật k-lân cận gần nhất
- cây quyết định
- Naïve Bayes

2. (**0.75 điểm**) a. Cho các xác xuất: P(C|S) = 0.5, P(C|~S) = 0.2, P(S) = 0.5. Hãy tính xác xuất có điều kiện P(S|C). (*0.75 điểm*)
Hint : Áp dụng công thức

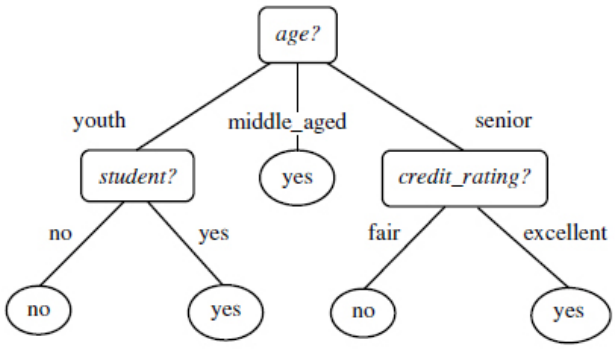
$$P(F_i \mid E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E \mid F_i)P(F_i)}{\sum_j P(E \mid F_j)P(F_j)}$$

3. (0.75 điểm) Nêu sự khác biệt giữa lựa chọn đặc trưng (feature selection) và trích yếu đặc trưng (feature abstraction). PCA là phương pháp lựa chọn đặc trưng hay trích yếu đặc trưng ?

4. (1. điểm) Chúng ta dùng bộ phân lớp *5-lân cận gần nhất có trọng số* (weighted 5-NN classifier) để phân lớp mẫu thử P. Giả sử khoảng cách giữa P với năm lân cận gần nhất (X_1, X_2, X_3, X_4 và X_5) lần lượt là $d_1 = 0.83$, $d_2 = 1.0$, $d_3 = 1.02$, $d_4 = 1.06$ và $d_5 = 1.08$. Nếu X_1, X4 thuộc lớp 3 và X_2, X_3 , X_5 thuộc lớp 2. Vậy P sẽ được phân vào lớp nào?

5. (0.75 điểm)

Cho một cây quyết định như ở hình bên, hãy suy dẫn ra tất cả các luật quyết định có dạng **if**..**then**… từ cây quyết định.



6. **(1.5 điểm)** Cho tập mẫu như sau:

Đặc trưng 1	Đặc trưng 2	Đặc trưng 3	Lớp
0	0	0	0
1	0	1	1
1	0	0	0
1	1	1	1
0	1	1	1
0	1	1	0

trong đó mỗi mẫu gồm 3 đặc trưng và nhãn lớp.
Nếu mẫu thử *P* với đặc trưng 1 là 0 và đặc trưng 2 là 0 và đặc trưng 3 là 1, hãy phân lớp mẫu thử này dùng giải thuật phân lớp Naïve Bayes.

7. (0.5 điểm) Cho ma trận đúng sai (confusion matrix) của một bộ phân lớp gồm 3 lớp như sau :

		Predicted		
		C_1	C_2	C_3
Actual	C_1	19	4	1
	C_2	3	20	4
	C_3	2	3	21

Tính độ chính xác (accuracy) của bộ phân lớp này.

8. (1.5 điểm)
a. Hãy nêu độ phức tạp tính toán của giải thuật k-means và giải thuật gom cụm phân cấp gộp. (*0.5 điểm*)
b. Nêu một thí dụ ứng dụng mà trong đó gom cụm được sử dụng như là một bước tiền xử lý (thu giảm tập huấn luyện) cho công tác phân lớp. (*0.5 điểm*)
c. Nếu tập mẫu có số mẫu là 9 và cần thiết phải tách tập mẫu này thành hai tập con phân ly. Như vậy có tổng cọng bao nhiêu cách tách có thể có? (*0.5 điểm*)

9. (1.5 điểm)
a) Tìm medoid của tập mẫu sau đây:
(1, 1), (1, 3), (1, 4), (2, 2), (2, 3), (3,1) (*0.75 điểm*)
b. Cho hai mẫu (mỗi mẫu gồm 8 thuộc tính nhị phân) X = (1, 0, 1, 1, 1, 0, 1, 1), Y = (0, 1, 1, 0, 0, 1, 0, 0), hãy tính khoảng cách giữa hai mẫu X và Y. (*0.25 điểm*)
c) Nêu điểm tương đồng giữa giải thuật gom cụm k-means và giải thuật gom cụm fuzzy-c-means. (*0.5 điểm*)

10. (1 điểm) Cho tập mẫu gồm 7 mẫu như sau: A = (1, 1), B = (1, 2), C = (2, 2), D = (6, 2), E = (7, 2), F = (6, 6), G = (7, 6).
Hãy mô tả diễn tiến của giải thuật k-means với k = 3 để gom cụm tập mẫu nêu trên, giả sử ba trung tâm cụm khởi đầu được chọn là A, D và F.