Đề kiểm tra giữa học kỳ I (2021-2022)
Môn : Học Máy và Ứng Dụng
Lớp Cao học
(Hạn chót nộp bài làm: 11:00 giờ PM ngày 24/10/2021
Địa chỉ email nộp bài: dtanhcse@gmail.com
Bài làm phải là: 1 file word hoặc 1 file PDF)

Đề kiểm tra gồm 2 trang

1. (0.75 điểm) Nêu sự khác biệt giữa *lựa chọn prototype* (prototype selection) và *trích yếu prototype* (prototype abstraction) khi thu giảm tập dữ liệu. Giải thuật Condensed Nearest Neighbors sử dụng phương pháp lựa chọn prototype hay trích yếu prototype ?

2. (1 điểm)
a. Cho một bộ phân lớp với 3 lớp C_1, C_2 và C_3. Hãy lập công thức tính độ đo accuracy của bộ phân lớp này. (*0.5 điểm*)
b. Confusion matrix của một bộ phân lớp với 3 lớp được cho 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

Hãy tính độ đo accuracy của bộ phân lớp này. (*0.5 điểm*)

3. (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 $d_1 = 1$, $d_2 = 3$, $d_3 = 4$, $d_4 = 5$ và $d_5 = 8$. Nếu X_1, X_2 thuộc lớp + và X_3, X_4 , X_5 thuộc lớp -. Vậy P sẽ được phân vào lớp nào?

4. (1.5 điểm) Cho một tập các mẫu hai chiều sau đây:
(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),
(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)
(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)
trong đó mỗi mẫu gồm đặc trưng thứ nhất, đặc trưng thứ hai và nhãn lớp.
Tìm hai điểm centroid của hai lớp 1 và 2. Sử dụng *minimum-distance classifier* để tìm lớp của mẫu thử *P* là (3.8, 3.1).

5. (0.75 điểm)
a. Cho các xác xuất: P(A|B) = 2/3, P(A|~B) = 1/3, P(B) = 1/3. Hãy tính xác xuất có điều kiện P(B|A).
Hint : Áp dụng công thức

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

b. Để có thể áp dụng bộ phân lớp Naive Bayes, chúng ta cần một giả định gì về tính chất dữ liệu.

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. **(1 điểm)**
Tìm centroid và medoid của tập mẫu sau đây:
(1, 1), (1, 3), (1, 4), (2, 2), (2, 3), (3,1)

8. (**0.5 điểm**) Nếu tập mẫu có số mẫu là 10 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?

9. (**0.5 điểm**) Nêu sự khác biệt giữa gom cụm cứng (hard clustering) và gom cụm mềm (soft clustering). Nêu điểm tương đồng giữa giải thuật gom cụm k-means và giải thuật gom cụm fuzzy-c-means.

10. (1 điểm)
a. Cho hai mẫu (mỗi mẫu gồm 3 thuộc tính) X = (7, 4, 3), Y = (4, 1, 8), hãy tính khoảng cách Manhattan giữa hai mẫu X và Y. (*0.5 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. (*0.5 điểm*)

11. (0.5 điểm)
Gom cụm gia tăng là gì? Nêu một nhược điểm của giải thuật gom cụm gia tăng Leader.