Ñeà thi giöõa hoïc kyø I (2020)
Moân : Học Máy và Ứng dụng
Lớp Cao học
Thôøi gian: 90 phuùt
(Sinh viên được phép tham khảo slides bài giảng)

Đề thi gồm 2 trang

1. **(1.5 điểm)** Trả lời các câu hỏi sau đây, mỗi câu 0.25 điểm:
1.1 (Đúng/Sai) Nếu P(A|B) = P(A) thì P(A,B) = P(A).P(B).
1.2 (Đúng/Sai) Giải thuật phân lớp Cây Quyết Định thì phù hợp với dữ liệu liên tục hơn là dữ liệu rời rạc (categorical data).
1.3 (Đúng/Sai) Khi giải thuật phân lớp Cây Quyết Định xây dựng cây quyết định với chiều sâu quá lớn, giải thuật có thể bị hiện tượng quá khớp (overfitting).
1.4 (Đúng/Sai) Để lựa chọn thuộc tính tách trong quá trình xây dựng Cây Quyết Định, chúng ta dựa vào việc cực tiểu hóa *mức suy giảm độ pha tạp* (drop in impurity)
1.5 Ma trận hiệp phương sai (covariance matrix) được sử dụng trong
A. Độ đo khoảng cách Mahalanobis
B. phương pháp PCA
C. Mô hình Gauss đa biến
D. Các trường hợp trên đều đúng
1.6 Đại diện cụm có thể là
A. centroid
B. medoid
C. leader
D. Các trường hợp trên đều đúng

2. (0.5 điểm) Nêu sự khác biệt giữa mô hình filter và mô hình wrapper trong công tác lựa chọn đặc trưng (feature selection).

3. (0.75 điểm) Nêu sự khác biệt giữa *lựa chọn prototype* (prototype selection) và *trích yếu prototype* (prototype abstraction). Giải thuật Condensed Nearest Neighbors sử dụng phương pháp lựa chọn prototype hay trích yếu prototype ?

4. (0.75 điểm) Nêu sự khác biệt giữa cách tiếp cận *maximum likelihood* (ML) (sử dụng trong giải thuật EM) và cách tiếp cận *maximum a posterior* (MAP) (sử dụng trong phương pháp Naïve Bayes).
6. (1.5 điểm) Cho một tập dữ liệu như sau, mỗi mẫu gồm 3 thuộc tính và một nhãn lớp:

Thuộc tính 1	Thuộc tính 2	Thuộc tính 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

Vì các thuộc tính của tập dữ liệu trên không là dữ liệu liên tục, nên chúng ta áp dụng phương pháp sau đây để tính khoảng cách giữa hai mẫu gồm các thuộc tính rời rạc.. Cho hai mẫu X và Y gồm *m* thuộc tính rời rạc, khoảng cách giữa *X* và *Y* là tổng số sự khác biệt giữa các trị thuộc tính tương ứng giữa hai mẫu. Nếu tổng số sự khác biệt càng nhỏ thì hai mẫu càng tương tự nhau. Công thức tính khoảng cách giữa X và Y như sau:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

với

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

Sử dụng độ đo khoảng cách nêu trên và phương pháp 1-lân cận gần nhất để phân lớp mẫu thử $P = (0, 0, 1)$, dựa vào tập huấn luyện là tập dữ liệu được cho ở trên.

7. (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?

8. (0.75 điểm) Giả sử khi trả lời một câu hỏi trong buổi thi trắc nghiệm, xác xuất để một thí sinh biết câu trả lời và trả lời đúng là 1/2 và xác xuất để thí sinh phải phỏng đoán câu trả lời là 1/2. Giả sử xác xuất để thí sinh trả lời đúng trong trường hợp biết câu trả lời là 1 và xác xuất để thí sinh trả lời đúng khi chỉ có thể phỏng đoán câu trả lời là 1/5 (vì mỗi câu hỏi trắc nghiệm có 5 lựa chọn). Vậy xác xuất có điều kiện để thí sinh biết câu trả lời trong trường hợp thí sinh đã trả lời đúng câu hỏi đó là bao nhiêu?
Hint: Để tính xác xuất $P(know\ answer \mid correct)$, áp dụng công thức Bayes.

9. **(1.5 điểm)** Cho một tập huấn luyện mà mỗi mẫu gồm 3 thuộc tính và được gắn nhãn lớp 1 hoặc 0 như sau :

Mẫu	A	B	C	Lớp
1	0	0	1	0
2	0	1	0	0
3	1	1	0	0
4	0	0	1	1
5	1	1	1	1
6	1	0	0	1
7	1	1	0	1

a) Ước lượng các xác xuất có điều kiện P(A=0|y=0), P(A=0|y=1), P(B=0|y=0), P(B=0|y=1), P(C=1|y=0), P(C=1|y=1).
b. Dựa vào những xác xuất được ước lượng ở câu a) hãy phân lớp mẫu thử (A = 0, B=0, C=1) dùng phương pháp phân lớp Naive Bayes.
10. (1 điểm) Giải thích hai phương pháp kiểm tra chéo (cross-validation) sau đây: kiểm tra chéo k-phần (k-fold cross validation) và kiểm tra chép bỏ ra một phần tử (leave-one-out cross validation).
11. (0.75 điểm) Chất lượng gom cụm của giải thuật k-means tùy thuộc vào những yếu tố nào? Giải thuật k-means phù hợp với dữ liệu liên tục hay dữ liệu rời rạc?