**Questions and Exercises**

**Course: Machine Learning**

### Chater 4 (Bayes Classifier)

**4.1** A study of a university found out that 15% of undergraduate students who smoke and 23% of graduate students who smoke. If 1/5 students in the university are graduate students and the rest are undergraduate students, what is the probability that a graduate student who smokes?

*Một nghiên cứu của một trường đại học cho thấy 15% sinh viên đại học hút thuốc và 23% sinh viên sau đại học hút thuốc. Nếu 1/5 sinh viên trong trường đại học là sinh viên cao học và số còn lại là sinh viên đại học, xác suất để một sinh viên cao học hút thuốc là bao nhiêu?*

**4.2** (True/false) If P(A|B) = P(A) then P(A,B) = P(A).P(B).

**4.3** State the difference between k-nearest neighbor algorithm and Naïve Bayes in classification.

*(Nêu sự khác biệt giữa thuật toán k-lân cận gần nhất và Naïve Bayes trong phân loại)*

**4.4** State the assumption on the characteristic of the dataset which allows us to apply Naïve Bayes classifier.

*(Nêu giả định về đặc điểm của tập dữ liệu cho phép chúng tôi áp dụng bộ phân loại Naïve Bayes.)*

**4.5** Consider the following data set: *(Hãy xem xét tập dữ liệu sau)*

| Feature 1 | Feature 2 | Feature 3 | Class |
| --- | --- | --- | --- |
| 0  1  1  1  0  0 | 0  0  0  1  1  1 | 0  1  0  1  1  1 | 0  1  0  1  1  0 |

If we have a test pattern *P* with feature 1 as 0 and feature 2 as 0 and feature 3 as 1, classify this pattern using Naïve Bayes classifier.

*(Nếu chúng ta có mẫu thử nghiệm P với đặc điểm 1 là 0 và đặc điểm 2 là 0 và đặc điểm 3 là 1, hãy phân loại mẫu này bằng bộ phân loại Naïve Bayes.)*

**4.6**. Given a dataset as in Exercise 4.4. Since the attributes are not continuous, we apply the following method to calculate the distance between two patterns with categorical attributes. Given two patterns, each consists of *m* categorical attributes. The distance between *X* and *Y* is total number of differences between the corresponding attribute values of the two patterns. The total number of differences is smaller, the two patterns more similar. That means:

*Cho một tập dữ liệu như trong Bài tập 4.4. Vì các thuộc tính không liên tục, nên chúng tôi áp dụng phương pháp sau để tính khoảng cách giữa hai mẫu có thuộc tính phân loại. Cho hai mẫu, mỗi mẫu bao gồm m thuộc tính phân loại. Khoảng cách giữa X và Y là tổng số khác biệt giữa các giá trị thuộc tính tương ứng của hai mẫu. Tổng số điểm khác biệt càng nhỏ thì hai mẫu càng giống nhau. Điều đó có nghĩa là:*



where



By using this distance measure, apply 1-nearest neighbor algorithm to classify the test pattern *P* = (0, 0, 1), based on the dataset given in Exercise 4.4.

Compare the results of the two classification method: 1-nearest neighbor algorithm and Naïve Bayes (Exercise 4.4).

*Bằng cách sử dụng thước đo khoảng cách này, hãy áp dụng thuật toán 1-láng giềng gần nhất để phân loại mẫu kiểm tra P = (0, 0, 1), dựa trên bộ dữ liệu được đưa ra trong Bài tập 4.4.*

*So sánh kết quả của hai phương pháp phân loại: thuật toán 1-láng giềng gần nhất và Naïve Bayes (Bài tập 4.4).*

**Chapter 5**

**(Decision Trees)**

**5.1.** Determine the entropy impurity for the following distributions.

a) The dataset has 1/2 of patterns belonging to the first class, 1/4 of patterns belonging to the second class, 1/8 of patterns belonging to the third class, 1/16 of patterns belonging to the fourth class, and 1/16 of patterns belonging to the fifth class.

b) The dataset consists of five classes and each class has 1/5 of patterns.

*Xác định tạp chất entropy cho các phân phối sau.*

*a) Tập dữ liệu có 1/2 số mẫu thuộc loại thứ nhất, 1/4 số mẫu thuộc loại thứ hai, 1/8 số mẫu thuộc loại thứ ba, 1/16 số mẫu thuộc loại thứ tư và 1/16 số mẫu thuộc lớp thứ năm.*

*b) Tập dữ liệu gồm 5 lớp và mỗi lớp có 1/5 số mẫu.*

**5.2.** Consider the following data set for a binary classification problem. Each pattern has two binary attributes and one class label (+ or -).

*Hãy xem xét tập dữ liệu sau đây cho một vấn đề phân loại nhị phân. Mỗi mẫu có hai thuộc tính nhị phân và một nhãn lớp (+ hoặc -).*

| A | B | Class label |
| --- | --- | --- |
| T | F | + |
| T | T | + |
| T | T | + |
| T | F | - |
| T | T | + |
| F | F | - |
| F | F | - |
| F | F | - |
| T | T | - |
| T | F | - |

Let use the information gain when determining the splitting attribute. Which of the features is selected as splitting attribute at the root node in the decision tree for the data set.

*(Hãy sử dụng thông tin thu được khi xác định thuộc tính phân tách. Tính năng nào được chọn làm thuộc tính phân tách tại nút gốc trong cây quyết định cho tập dữ liệu.)*

**5.3.** Consider the following Weather dataset for a binary classification problem. Each pattern has four discrete attributes and one class label (Yes or No).

*Hãy xem xét tập dữ liệu Thời tiết sau đây cho một vấn đề phân loại nhị phân. Mỗi mẫu có bốn thuộc tính rời rạc và một nhãn lớp (Có hoặc Không).*

| Outlook | Temperature | Humidity | Windy | Play Tenis |
| --- | --- | --- | --- | --- |
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | False | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | High | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Sunny | Warm | High | False | No |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Warm | Normal | False | Yes |
| Sunny | Warm | Normal | True | Yes |
| Overcast | Warm | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Rainy | Warm | High | True | No |

Let use the information gain when determining the splitting attribute. Which of the features is selected as splitting attribute at the root node in the decision tree for the data set.

*(Hãy sử dụng thông tin thu được khi xác định thuộc tính phân tách. Tính năng nào được chọn làm thuộc tính phân tách tại nút gốc trong cây quyết định cho tập dữ liệu.)*

**5.4** (True/false) The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

*(Đúng/sai) Độ sâu của cây quyết định đã học có thể lớn hơn số ví dụ huấn luyện được sử dụng để tạo cây.*

**Chương 6**

**(Clustering)**

**6.1** State the difference between supervised learning (classification) and unsupervised learning (clustering).

*Nêu sự khác biệt giữa học có giám sát (phân loại) và học không giám sát (phân cụm).*

**6.2** Consider the following 10 patterns:

X1 = (1, 1), X2 = (6, 1), X3 = (2, 1), X4 = (6, 7), X5 = (1, 2), X6 = (7, 1), X7 = (7, 7), X8 = (2, 2), X1 = (6, 2), X10 = (7, 6)

Obtain the distance matrix using the Euclidean distance as the distance between two patterns.

*Lấy ma trận khoảng cách sử dụng khoảng cách Euclide là khoảng cách giữa hai mẫu.*

**6.3** If there is a set of *n* patterns and it is required to cluster these patterns to form two clusters, how many such partitions will there be?

*Nếu có một tập hợp n mẫu và cần phải phân cụm các mẫu này để tạo thành hai cụm, thì sẽ có bao nhiêu phân vùng như vậy?*

**6.4.** Given the cluster of 5 patterns:

X1 = (1, 1), X2 = (1, 2), X3 = (2, 1), X4 = (1.6, 1.4), X5 = (2, 2)

Show that the medoid of the cluster is (1.6, 1.4).

*Chứng tỏ rằng trung điểm của cụm là (1.6, 1.4).*

**6.5** In agglomerative hierarchical clustering, among the current clusters, how to select the most suitable pair of clusters to be merged?

*Trong phân cụm theo thứ bậc tập hợp, trong số các cụm hiện tại, làm thế nào để chọn ra cặp cụm phù hợp nhất để hợp nhất?*

**6.6** In divisive hierarchical clustering,

a. How to find the best way to split a cluster into two clusters.

b. How to select among the current clusters the most suitable cluster to be split.

*Trong phân cụm phân cấp phân chia,*

*a. Cách tìm cách tốt nhất để tách một cụm thành hai cụm.*

*b. Làm thế nào để chọn trong số các cụm hiện tại cụm phù hợp nhất để tách.*

**6.7** State the computational complexity of k-means algorithm.

*Nêu độ phức tạp tính toán của thuật toán k-means.*

**6.8** State the strong points and weak points of k-means algorithm.

*Nêu điểm mạnh và điểm yếu của thuật toán k-means.*

**6.9**. State the computational complexity of agglomerative hierarchical clustering and divisive hierarchical clustering.

*Nêu độ phức tạp tính toán của phân cụm theo thứ bậc tổng hợp và phân cụm theo thứ bậc phân chia.*

**6.10** Consider the two dimensional data set given below:

(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)

Use the k-means algorithm to cluster these patterns with k = 3.

*Hãy xem xét tập dữ liệu hai chiều được đưa ra dưới đây:*

*(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5 , 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)*

*Sử dụng thuật toán k-means để nhóm các mẫu này với k = 3.*

**6.11** Given a set of 2-dimensional patterns: X1 = (1, 3), X2 = (1.5, 3.2), X3 = (1.3, 2.8), X4 =(3,1). Let apply k-means with k=2 to cluster this dataset. Assume that at a certain iteration, the dataset is grouped into 2 clusters as follows : the first cluster consists of X1, and the second cluster consists of X2, X3, X4. Let perform the next iteration which consists of two steps : recalculating the centroids and assigning the patterns to the clussters.

Note: Euclidean distance is used in the k-means.

*Cho một tập hợp các mẫu 2 chiều: X1 = (1, 3), X2 = (1,5, 3,2), X3 = (1,3, 2,8), X4 =(3,1). Hãy áp dụng phương tiện k với k=2 để phân cụm tập dữ liệu này. Giả sử rằng tại một lần lặp nhất định, tập dữ liệu được nhóm thành 2 cụm như sau: cụm thứ nhất gồm X1 và cụm thứ hai gồm X2, X3, X4. Hãy thực hiện lần lặp tiếp theo bao gồm hai các bước: tính toán lại các trọng tâm và gán các mẫu cho các cụm.*

***Lưu ý:*** *Khoảng cách Euclide được sử dụng trong k-mean.*

**6.12** Consider the two dimensional data set given below:

(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)

Use agglomerative hierarchical clustering with single-link and agglomerative hierarchical clustering with complete-link to cluster the dataset into 4 clusters.

*Hãy xem xét tập dữ liệu hai chiều được đưa ra dưới đây:*

*(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5 , 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)*

*Sử dụng phân cụm theo cấp bậc tổng hợp với liên kết đơn và phân cụm theo cấp bậc tổng hợp với liên kết đầy đủ để phân cụm tập dữ liệu thành 4 cụm.*

**6.13** State the similarity between two clustering algorithms: k-means and fuzzy-c-means.

*Nêu sự giống nhau giữa hai thuật toán phân cụm: k-mean và Fuzzy-c-mean.*

**6.14** Given a set of 2-dimensional patterns: X1 = (1, 6), X2 = (2,5), X3 = (3, 8), X4 =(4,4). X5 = (5, 7), X6 =(6,9). Let apply fuzzy-c-means with k = 2 to cluster this dataset. Assume that at a certain iteration, the dataset is grouped into 2 clusters with the membership weights as follows.

|  | X1 | X2 | X3 | X4 | X5 | X6 |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster c1 | 0.8 | 0.9 | 0.7 | 0.3 | 0.5 | 0.2 |
| Cluster c2 | 0.2 | 0.1 | 0.3 | 0.7 | 0.5 | 0.8 |

Let perform the next iteration which consists of two steps : recalculating the centroids and assigning the membership weights for each pattern.

Note: Euclidean distance is used in the fuzzy-c-means.

*Cho một tập các mẫu 2 chiều: X1 = (1, 6), X2 = (2,5), X3 = (3, 8), X4 =(4,4). X5 = (5, 7), X6 =(6,9). Hãy áp dụng Fuzzy-c-means với k = 2 để phân cụm tập dữ liệu này. Giả sử rằng tại một lần lặp nhất định, tập dữ liệu được nhóm thành 2 cụm với trọng số thành viên như sau.*

*…*

*Hãy thực hiện bước lặp tiếp theo bao gồm hai bước: tính toán lại trọng tâm và gán trọng số thành viên cho mỗi mẫu.*

*Lưu ý: Khoảng cách Euclide được sử dụng trong phương pháp fuzzy-c-means.*

**6.15**. (True/False) K-Means can generate clusters with arbitrary shapes.

*K-Means có thể tạo các cụm có hình dạng tùy ý.*

**6.16**. (True/False) DBSCAN can generate clusters with arbitrary shapes.

*DBSCAN có thể tạo các cụm có hình dạng tùy ý*

**6.17**. (True/False) K-Means can generate clusters only with spherical shapes.

*K-Means chỉ có thể tạo các cụm có hình cầu.*

**6.18** Give an example in which clustering can be used as a preprocessing step for an another data classification task.

*Cho một ví dụ trong đó phân cụm có thể được sử dụng như một bước tiền xử lý cho một nhiệm vụ phân loại dữ liệu khác.*

**6.19** Explain the term *incremental clustering*. State the weak point of the Leader algorithm for incremental clustering.

*Giải thích thuật ngữ phân cụm gia tăng. Nêu điểm yếu của thuật toán Leader để phân cụm tăng dần.*

**6.20** Given a set of 2-dimensional patterns :

A = (1, 1), B = (1, 2), C = (2, 2), D = (6, 2), E = (7, 2), F = (6, 6), G = (7, 6)

Let apply Leader algorithm to cluster the dataset. Assume that the data will be processed in the order A, B, C, D, E, F and G, and the user specified threshold *T* be 3.

*Cho một tập hợp các mẫu 2 chiều:*

*A = (1, 1), B = (1, 2), C = (2, 2), D = (6, 2), E = (7, 2), F = (6, 6), G = (7, 6)*

*Hãy áp dụng thuật toán Leader để phân cụm tập dữ liệu. Giả sử rằng dữ liệu sẽ được xử lý theo thứ tự A, B, C, D, E, F và G và ngưỡng T do người dùng chỉ định là 3.*

**6.21** How to evaluate clustering quality based on objective function.

*Cách đánh giá chất lượng phân cụm dựa trên hàm mục tiêu.*