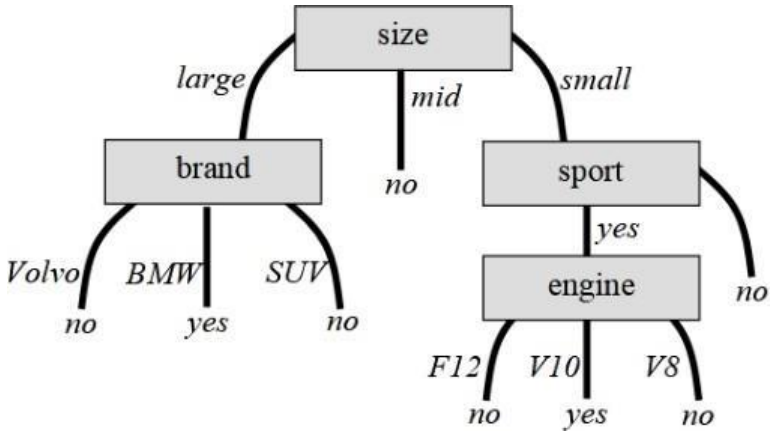
MÁY HỌC ỨNG DỤNG

Câu 1:

Cho cây quyết định như hình bên dưới, NHỮNG phát biểu nào sau đây là phát biểu SAI:



a. Thuộc tính của tập dữ liệu: size, brand, sport, engine
b. Thuộc tính của tập dữ liệu: large, mid, small, yes, no
c. Thuộc tính của tập dữ liệu: size, brand, sport, engine, yes, no
d. Thuộc tính của tập dữ liệu: F12, V10, V8, Volvo, BMW, SUV

1.(a), (b) và (c)
2.(a), (b) và (d)
3.(a), (c) và (d)
4.(b), (c) và (d)

Câu 2:

Cho tập dữ liệu như bảng bên dưới, người ta cần huấn luyện mô hình để khi có một phần tử mới có các thông tin sau: "X1=35", "X2=32", "X3=1" thì người ta có thể dự đoán được giá trị "Y" của nó, anh chị cần lựa chọn giải thuật nào sau đây để thực hiện yêu cầu trên?

Y	X1	X2	X3
228	67	36	1
93	4	16	0
186	81	29	1
106	82	27	0
75	55	29	0
83	39	26	1
57	14	30	1
74	33	45	0
93	71	34	0

a. Bayes thơ ngây (Naïve Bayes)
b. Kmeans
c. Hồi quy tuyến tính (regression)
d. Cả ba đều đúng

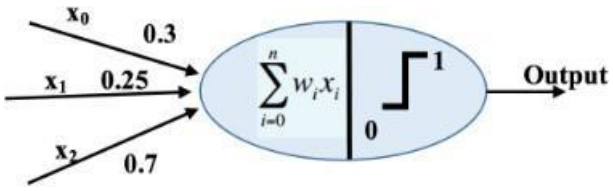
Câu 3:

Cho tập dữ liệu chứa các thông tin về khách hàng của một ngân hàng: tuổi, nghề nghiệp, số lượng thẻ tín dụng, thu nhập hàng tháng. Ngân hàng cần tìm ra các nhóm khách hàng khác nhau để có những gói dịch vụ phù hợp cho các nhóm đối tượng này. Chúng ta cần sử dụng giải thuật học gì để giúp các nhà quản lý của ngân hàng này?

a. Giải thuật học không có giám sát, giải thuật phân lớp.
b. Giải thuật học có giám sát, giải thuật hồi quy.
c. Giải thuật học có giám sát, giải thuật phân lớp.
d. Giải thuật học không có giám sát, giải thuật gom nhóm/gom cụm.

Câu 4:

Cho "perceptron" như hình, phát biểu nào sau đây là phát biểu ĐÚNG?



a. Các giá trị 0.3, 0.25 và 0.7 là giá trị trọng số tương ứng của các thuộc tính, hàm truyền là một hàm ngưỡng [0,1]
b. Đây là perceptron tuyến tính có ngưỡng với hàm kích hoạt là một hàm ngưỡng [0,1]
c. Đây là perceptron tuyến tính có ngưỡng với hàm truyền là một hàm ngưỡng [0,1]
d. Các giá trị 0.3, 0.25 và 0.7 là giá trị các thuộc tính, hàm kích hoạt là một hàm ngưỡng [0,1].

Câu 5:

Giả sử tập dữ liệu có 1.000 mẫu tin (là các tin nhắn) trong đó có 10 tin nhắn rác và 990 tin nhắn bình thường. Có hai mô hình phân lớp M1 và M2 cho kết quả tương ứng như sau:

Ma trận confusion thu được từ mô hình M1

		Dự đoán	
		Tin nhắn rác	Tin nhắn bình thường
Thực tế	Tin nhắn rác	1	9
	Tin nhắn bình thường	10	980

Ma trận confusion thu được từ mô hình M2

		Dự đoán	
		Tin nhắn rác	Tin nhắn bình thường
Thực tế	Tin nhắn rác	10	0
	Tin nhắn bình thường	20	970

Anh/chị hãy cho biết mô hình nào thích hợp để xử lý tập dữ liệu trên? Giải thích lý do tại sao?

a. Mô hình M1 vì chỉ số F1 của mô hình M1 nhỏ hơn chỉ số F1 của mô hình M2
b. Mô hình M2 vì chỉ số F1 của mô hình M2 lớn hơn chỉ số F1 của mô hình M1
c. Mô hình M1 vì chỉ số F1 của mô hình M1 lớn hơn chỉ số F1 của mô hình M2
d. Mô hình M2 vì chỉ số F1 của mô hình M2 nhỏ hơn chỉ số F1 của mô hình M1

Câu 6:

Các phát biểu nào sau đây là phát biểu đúng?

a. Giải thuật KNN không có quá trình học nên khi phân loại có tốc độ nhanh hơn các giải thuật khác
b. Giải thuật KNN không có quá trình học nên khi phân loại có tốc độ chậm hơn các giải thuật khác
c. Giải thuật KNN giải quyết cho bài toán phân loại và bài toán gom nhóm

Câu 7:

Chúng ta xem xét một mô hình nhận dạng hoa hồng và hoa cúc. Tập dữ liệu gồm có 1000 hoa hồng và 950 hoa cúc, được chia thành 2 phần là tập huấn luyện và tập kiểm thử với tỷ lệ 70/30. Sau khi huấn luyện, độ chính xác của mô hình đạt được 30% trên tập huấn luyện và 20% trên tập kiểm tra. Theo anh chị mô hình này như thế nào?

a. Chưa khớp (Underfitting)
b. Quá khớp (Overfitting)
c. Vừa khớp (Good Fitting)

Câu 8:

Cho tập dữ liệu như sau (dữ liệu không bao gồm cột STT), người ta cần chia dữ liệu này thành 2 nhóm, anh chị cần lựa chọn giải thuật nào sau đây để thực hiện yêu cầu trên?

STT	Ever_married	Age	Graduated	Work Experience
1	No	26	Yes	A
2	No	19	No	D
3	No	17	No	D
4	Yes	70	No	A
5	Yes	58	No	B
6	No	41	No	C
7	No	32	No	D
8	No	31	No	B

a. Cây quyết định - Decision tree
b. K láng giềng - KNN
c. Bayes thơ ngây - Naive Bayes
d. Kmeans

Câu 9:

Phát biểu nào sau đây đúng cho "Boosting technique"?

a. Giúp giảm lỗi Variance
b. Xây dựng các mô hình độc lập nhau
c. Tập trung cải tiến lỗi sinh ra từ mô hình trước

Câu 10:

Phát biểu nào sau đây đúng cho "Bagging technique"?

a. Tập trung cải tiến lỗi sinh ra từ mô hình trước
b. Bagging có thể thực hiện song song
c. Giúp giảm lỗi Bias

Câu 11:

Cho tập dữ liệu gồm 5 phần tử với 3 thuộc tính BMI, huyết áp, Glucose và nhãn tương ứng như bảng bên dưới. Sử dụng giải thuật KNN với để dự đoán cho phần tử mới tới, chúng ta cần tính khoảng cách của phần tử mới tới đến tất cả các phần tử của tập dữ liệu. Anh/chị hãy cho biết khoảng cách từ phần tử mới tến đến phần tử thứ 3 (tô màu cam) là bao nhiêu nếu sử dụng khoảng cách Manhatan.

a. 65
b.53.75
c.2889
d.70

STT	BMI	Huyết áp	Glucose	Nhãn
1	33	72	148	Yes
2	26	6	85	*No*
3	25	67	183	*Yes*
4	28	66	89	No
5	30	74	145	Yes
Người dùng mới	29	75	130	

Câu 12:

Cho thông tin như sau: tuổi, tình trạng mang thai, chỉ số Glucose, huyết áp, chỉ số BMI, tình trạng bệnh tiểu đường (có/không). Theo anh/chị, chúng ta nên sử dụng giải thuật học gì để dự báo bệnh nhân có bị bệnh tiểu đường hay không?

a. Giải thuật học có giám sát, giải thuật phân lớp.
b. Giải thuật học không có giám sát, giải thuật phân lớp.
c. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm.
d. Giải thuật học có giám sát, giải thuật hồi quy

Câu 13:

Chúng ta xem xét một mô hình phân loại có khả năng nhận dạng ảnh hoa hồng và hoa cúc. Tập kiểm tra gồm có 100 ảnh hoa hồng và 75 ảnh hoa cúc. Đối với 100 ảnh hoa hồng, mô hình nhận dạng đúng 80 ảnh. Đối với ảnh hoa cúc, mô hình nhận dạng đúng 60 ảnh. Hãy cho biết độ chính xác tổng thể (accuracy) của mô hình này?

a. (80+60) / (100+75)
b. 80/100
c. (80/100) + (60/75)
d. 60/75

Câu 15:

Hiện tượng overfitting (học vẹt) xảy ra trong tình huống nào sau đây?

a. Xảy ra khi sai số dự đoán của mô hình trên tập huấn luyện cao, nhưng trên tập kiểm thử thì thấp
b. Xảy ra khi sai số dự đoán của mô hình trên cả tập huấn luyện và tập kiểm thử đều thấp.
c. Cả a và b đều sai
d. Cả a và b đều đúng

Câu 16:

Cho tập dữ liệu như bảng bên dưới gồm 6 phần tử với 2 thuộc tính màu tóc, độ bóng mượt và nhãn là tình trạng tóc. Anh/chị hãy tính độ hỗn loạn thông tin sau khi phân hoạch cho thuộc tính “độ bóng mượt" với điểm phân hoạch là 54

STT	Màu tóc	Độ bóng mượt	Tình trạng tóc
1	Đen	40	**Bị cháy nắng**
2	Nâu	46	***Bị cháy nắng***
3	Đen	58	**Bị cháy nắng**
4	Đen	56	***Không cháy nắng***
5	Đen	53	***Không cháy nắng***
6	Nâu	61	***Không cháy nắng***

a.0.528
b.0.389
c.1
d.0.918

Câu 17:

Đặc trưng trong máy học là gì?
a. Hàm mục tiêu
b. Thuộc tính có giá trị quyết định đối với mục tiêu
c. Dữ liệu đầu vào
d. Phương pháp đánh giá mô hình

Câu 18:

Chúng ta xem xét một mô hình phân loại có khả năng nhận dạng ảnh con chó và con mèo. Tập kiểm tra gồm có 100 ảnh con chó và 80 ảnh con mèo. Trong 100 ảnh con chó, mô hình nhận dạng đúng 80 ảnh. Đối với ảnh con mèo, phần mềm nhận dạng đúng 60 ảnh. Hãy cho tính độ chính xác (accuracy) của mô hình này?

a. (80+60) / (100+80)
b. (80/100) + (60/80)
c. 60/80
d. 80/100

Câu 19:

Cho tập dữ liệu gồm 8 phần tử, mỗi phần tử có 2 thuộc tính: "trọng lượng", "màu sắc" như bảng bên dưới. Dựa vào 2 thuộc tính này người ta phân loại "chuẩn xuất khẩu" của thanh long: "đạt" hay "không đạt".

STT	Trọng lượng (gram)	Màu sắc	Chuẩn xuất khẩu
1.	300	Đỏ	**Đạt**
2.	280	Đỏ	**Không đạt**
3.	290	Xanh	**Không đạt**
4.	350	Đỏ	**Đạt**
5.	250	Xanh	**Không đạt**
6.	310	Xanh	**Đạt**
7.	320	Đỏ	**Đạt**
8.	270	Xanh	**Không đạt**

Người ta đã xây dựng được mô hình Bayes thơ ngây như bên dưới, anh/chị hãy xác định xác suất của phần tử A cho nhãn đạt và không đạt để dự đoán nhãn của thanh long A có trọng lượng = 299 và màu sắc là đỏ.(biết f(trọng lượng =299/đạt) ~ 0.0115 và f(trọng lượng =299/không đạt))= 0.00701

Màu sắc	Đạt	Không đạt
Xanh	0.25	0.75
Đỏ	0.75	**0.25**

Trọng lượng	Đạt	Không đạt
Mean	320	272.5
σ^2	**466.67**	291.67
σ	21.6	17.08

P(Đạt)	0.5	P(Không đạt)	0.5

a.P(đạt/A) = (0.0086/ P(A)); P (không đạt/A) = (0.0018/ P(A)): nhãn Không Đạt
b.P(đạt/A) = (0.0043/ P(A)); P (không đạt/A) = (0.00088/ P(A)): nhãn Đạt c.P(đạt/A) = (0.0086/ P(A)); P (không đạt/A) = (0.0018/ P(A)): nhãn Đạt d.P(đạt/A) = (0.0043/ P(A)); P (không đạt/A) = (0.00088/ P(A)): nhãn Không Đạt

Câu 20:

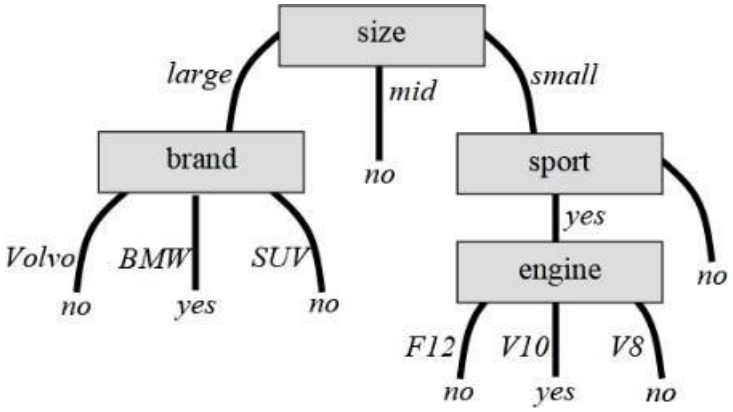
Cho tập dữ liệu gồm 10 phần tử. Mỗi phần tử gồm có 2 biến độc lập X1, X2 và 1 biến phụ thuộc Y. Tập dữ liệu này được chia thành 2 phần, tập huấn luyện và tập kiểm thử. Sau khi huấn luyện, ta tìm được giá trị theta0 = 0.3; theta1=0.46; theta2 = 0.32. Giả sử tập kiểm thử có giá trị như sau, hãy tính chỉ số MAE (Mean Absolute Error).

STT	X1	X2	Y
1	3	9	9
2	5	8	10
3	8	10	15

a.6
b.5.7
c.6.2
d.5.4

Câu 21:

Cho cây quyết định như hình bên dưới, NHỮNG phát biểu sau đây là SAI:



a. Nhãn của tập dữ liệu: size, brand, sport, engine
b. Nhãn của tập dữ liệu: size, brand, sport, engine, yes, no
c. Nhãn của tập dữ liệu: F12, V10, V8, Volvo, BMW, SUV
d. Nhãn của tập dữ liệu: yes, no
1. (b), (c) và (d
2. (a), (b) và (d)
3. (a), (b) và (c)
4. (a), (c) và (d)

Câu 22:

Cho tập dữ liệu gồm 10 phần tử. Mỗi phần tử gồm có 2 biến độc lập X1, X2 và 1 biến phụ thuộc Y. Tập dữ liệu này được chia thành 2 phần, tập huấn luyện và tập kiểm thử. Sau khi huấn luyện, ta tìm được giá trị theta0 = 0.3; theta1=0.46; theta2 = 0.32. Giả sử tập kiểm thử có giá trị như bảng bên dưới, hãy chọn chỉ số đánh giá phù hợp.

STT	X_1	X_2	Y
1	3	9	9
2	5	8	10
3	8	10	15

a. Precision
b. Recall
c. F1
d. Mean squared error (MSE)

Câu 23:

Trích xuất đặc trưng thường được thực hiện ở đâu trong quy trình xây dựng mô hình máy học?

a. Trước và sau quá trình xử lý dữ liệu
b. Khi thực hiện huấn luyện
c. Trong quá trình lựa chọn mô hình
d. Trước quá trình xử lý dữ liệu

Câu 24:

Phát biểu nào sau đây là đúng đối với averaging technique?

a. Chỉ được dùng cho bài toán phân lớp
b. Chỉ được dùng cho bài toán hồi quy
c. Được dùng cho cả bài toán phân lớp và bài toán hồi quy
d. Tất cả đều sai
d.6

Câu 25:

Phương pháp tập hợp mô hình (Ensemble Methods) là kết hợp nhiều mô hình cơ sở dựa trên tập học nhằm cải thiện độ chính xác của giải thuật dự đoán. Tăng hiệu quả của mô hình dựa trên cơ sở giảm lỗi bias, variance. Lỗi bias được hiểu như thế nào?

a. Lỗi bias là lỗi liên quan đến mô hình (bộ phân lớp/ dự đoán) mà không liên quan đến dữ liệu được dùng để huấn luyện
b. Lỗi Bias là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học (data samples).
c. Lỗi Bias là lỗi liên quan đến việc lựa chọn dữ liệu để huấn luyện mô hình
d. Tất cả đều đúng

Câu 26:

Để đánh giá hiệu quả của mô hình hồi quy tuyến tính, người ta có thể dùng những chỉ số nào sau đây?

(a) MAE
(b) Accuracy
(c) Recall
(d) F1
(e) RMSE

1.(a) và (e)
2.(a) và (d)
3.(a) và (c)
4.(a) và (b)

Câu 27:

Cho tập dữ liệu như bảng bên dưới gồm 6 phần tử với 2 thuộc tính màu sắc, trọng lượng và nhãn là chuẩn xuất khẩu. Anh/chị hãy xác định chỉ số Gini của thuộc tính "màu sắc"

STT	Trọng lượng (gram)	Màu sắc	Chuẩn xuất khẩu
1.	300	Xanh	**Đạt**
2.	280	Đỏ	**Không đạt**
3.	290	Xanh	**Không đạt**
4.	350	Đỏ	**Đạt**
5.	250	Xanh	**Không đạt**
6.	310	Xanh	**Đạt**

a.0.25
b.0.5
c.1
d.0.44

Câu 28:

Để đánh giá hiệu quả của bài toán phân lớp với số lượng lớp lớn hơn 2, ta sử dụng chỉ số nào sau đây?

a.MAE
b.F1
c.Accuracy
d.MSE hoặc RMSE

Câu 29:

Chúng ta xem xét một mô hình nhận dạng hoa hồng và hoa cúc. Tập dữ liệu gồm có 1000 hoa hồng và 950 hoa cúc, được chia thành 2 phần là tập huấn luyện và tập kiểm thử với tỷ lệ 70/30. Sau khi huấn luyện, độ chính xác của mô hình đạt được 95% trên tập huấn luyện và 20% trên tập kiểm tra. Theo anh chị mô hình này như thế nào?

a. Vừa khớp (Good Fitting)
b. Quá khớp (Overfitting)
c. Chưa khớp (Underfitting)

Câu 30:

Để đánh giá mô hình dự báo mật số rầy nâu gây hại trên lúa (được xây dựng bằng giải thuật hồi quy), bạn sẽ sử dụng tiêu chí nào sau đây.

a. Mean squared error (MSE)
b. Precision
c. F1
d. Recall

Câu 31:

Xem xét một nghiên cứu với tập dữ liệu có 9360 dòng. Mỗi dòng là một ngày và mỗi dòng có 6 giá trị là nhiệt độ (tmax, tmin), bức xạ mặt trời (solar radiation), hướng gió (wind-dir), tốc độ gió (wind-speed) và lượng mưa (rainfall). Tất cả các tham số này đều có giá trị thực. Anh/chị được giao nhiệm vụ xây dựng mô hình dự báo lượng mưa (rainfall) với đơn vị đo mm từ 5 thuộc tính còn lại. Bài toán trên thuộc dạng nào sau đây?

a. Bài toán gom cụm
b. Thỏa mãn ràng buộc
c. Bài toán phân lớp
d. Bài toán hồi quy

Câu 32:

Mô hình được gọi là underfitting khi:

a. Sai số dự đoán của mô hình trên tập huấn luyện cao, nhưng trên tập kiểm thử thì thấp
b. Sai số dự đoán của mô hình trên cả tập huấn luyện và tập kiểm thử đều thấp
c. Sai số dự đoán của mô hình trên tập huấn luyện thấp, nhưng trên tập kiểm thử thì cao
d. Sai số dự đoán của mô hình trên cả tập huấn luyện và tập kiểm thử đều cao.

Câu 33:

Để đánh giá hiệu quả của mô hình hồi quy tuyến tính, người ta dùng chỉ số nào sau đây?
a. Recall
b. Accuracy
c. F1
d. MSE

Câu 34:

Giải thuật cây quyết định được xây dựng Top-down, bắt đầu nút gốc, tất cả các dữ liệu học ở nút gốc. Nếu dữ liệu tại 1 nút có cùng lớp -> nút lá (nhãn của nút chính là nhãn của các phần tử thuộc nút lá); Nếu dữ liệu ở nút chứa các phần tử có lớp rất khác nhau (không thuần nhất) thì phân hoạch dữ liệu một cách đệ quy bằng "việc chọn 1 thuộc tính để thực hiện phân hoạch tốt nhất có thể" => kết quả thu được cây nhỏ nhất.

Việc chọn 1 thuộc tính để thực hiện phân hoạch tốt nhất có thể trong định nghĩa trên được hiểu như thế nào là đúng?

a. Chọn thuộc tính có độ lợi thông tin nhỏ nhất và chỉ số gini lớn nhất.
b. Chọn thuộc tính có độ lợi thông tin nhỏ nhất
c. Chọn thuộc tính có chỉ số gini lớn nhất
d. Chọn thuộc tính có độ lợi thông tin lớn nhất hoặc chỉ số gini nhỏ nhất.

Câu 35:

Cho bảng dữ liệu gồm 6 phần tử với các thông tin X1, X2, X3. Người ta đã tính khoảng cách từ phần tử mới tới là Hưng với các thuộc tính < X1=30, X2=50, X3=2> đến 6 phần tử có trong tập dữ liệu như cột "Khoảng cách". Theo anh/chị nhãn của Hưng là gì khi K=3, tại sao?

Phần tử	X1	X2	X3	Khoảng cách	Nhãn
Nam	35	35	3	***21***	***Không***
Lan	22	50	2	***8***	***Có***
Ngọc	25	40	4	***17***	***Có***
Mai	63	200	1	***184***	***Không***
Dũng	59	170	1	***150***	***Không***
Tuấn	37	50	2	*7*	***Không***

a. Nhãn là "Không" vì Hưng gần với Mai, Dũng và Nam
b. Nhãn là "Không" vì Hưng gần với Tuấn
c. Nhãn là "Không" vì Hưng gần với Mai nhất
d. Nhãn là "Có" vì Hưng gần với Lan, Ngọc và Tuấn

Câu 36:

Cho tập dữ liệu chứa các thông tin về rượu vang như sau: độ Ph, độ cồn, lượng đường, axit cố định, axit bay hơi, thời gian xuất xưởng (theo đơn vị giờ). Chúng ta cần sử dụng giải thuật học gì để giúp dự đoán thời gian xuất xưởng của rượu khi biết thông tin độ Ph, độ cồn, lượng đường, axit cố định, axit bay hơi của loại rượu đó?

a. Bài toán gom cụm (clustering)
b. Tìm kiếm thoả mãn ràng buộc
c. Bài toán hồi quy (regression)
d. Bài toán phân lớp (classification)

Câu 37:

Để xây dựng mô hình dự báo dịch rầy nâu gây hại trên lúa, các nhà nghiên cứu của một trường đại học đã thu thập dữ liệu tại địa bàn Trung An, Quận Thốt Nốt, Thành Phố Cần Thơ. Tập dữ liệu thu được là kết quả điều tra tại 840 địa điểm (phần tử), với 24 thuộc tính khác nhau. Sau khi tiền xử lý, loại bỏ các thuộc tính không dùng trong dự báo như: số thứ tự, mã ruộng, các thuộc tính có dữ liệu nhiễu và số liệu điều tra sai lệch cũng được bỏ qua như: ngày điều tra, ngày sạ, tuổi lúa. Các nhà nghiên cứu đã thu được 12 thuộc tính, trong đó có 11 thuộc tính dự báo dùng để xây dựng mô hình dự báo mật số rầy (thuộc tính phụ thuộc, có giá trị từ 0 đến 12900). Các thuộc tính dự báo bao gồm:

1. Kinh độ
2. Vĩ độ
3. Giống lúa
4. Mật độ sạ (kg/ha)
5. Nhiệt độ không khí (độ C)
6. Ẩm độ không khí (%)
7. Mực nước ruộng (cm)
8. So màu lá lúa (số màu: 1/2/3/4/5/6)
9. Mật số cỏ (cây/m2)
10. Số chồi/m2
11. Số lá/m2

Anh/chị chọn những chỉ số nào sau đây để đánh giá mô hình dự báo mật số rầy nâu từ 11 thuộc tính này.

(a) MAE
(b) Accuracy
(c) Recall
(d) F1
(e) RMSE

1.(a) và (b)
2.(a) và (c)
3.(a) và (e)
4.(a) và (d)

Câu 38:

Trong một cuộc bầu cử, N người bầu cử đang bỏ phiếu cho một trong hai ứng cử viên. Những người bỏ phiếu không giao tiếp với nhau trong khi bỏ phiếu. Phương pháp tập hợp mô hình nào sau đây hoạt động tương tự như quy trình bầu cử vừa nêu trên?

a. Bagging
b. Boosting
c. Cả a và b đều đúng

Câu 39:

Cho tập dữ liệu gồm 5 phần tử với 3 thuộc tính thời gian sử dụng dịch vụ, số tiền sử dụng, số lượng dịch vụ. Sử dụng giải thuật KNN để dự đoán cho phần tử mới tới, chúng ta cần tính khoảng cách của phần tử mới tới đến tất cả các phần tử của tập dữ liệu. Anh/chị hãy cho biết khoảng cách từ phần tử mới có giá trị thời gian sử dụng dịch vụ=7, số tiền sử dụng=4, số lượng dịch vụ=2 đến phần tử thứ 5 (tô màu cam) là bao nhiêu nếu sử dụng khoảng cách Manhatan.

a.2.4
b.1.4
c.2

Thời gian sử dụng dịch vụ	Số tiền sử dụng(triệu)	Số lượng dịch vụ	Nhãn
9	4	2	B
5	3	1	A
8	2	3	B
7	4	2	B
6	5	2	B

Câu 40:

Cho tập dữ liệu gồm 15 phần tử. Mỗi phần tử gồm có 2 biến độc lập X1, X2 và 1 biến phụ thuộc Y. Tập dữ liệu này được chia thành 2 phần, tập huấn luyện và tập kiểm thử. Sau khi huấn luyện, ta tìm được giá trị theta0 = 0.3; theta1 = 0.66; theta2 = 0.72. Giả sử tập kiểm thử có giá trị như sau, hãy tính chỉ số MAE (Mean Absolute Error).

STT	X1	X2	Y
1	3	9	9
2	5	8	10.04
3	8	12	15
4	4	8	9
5	6	7	9.1

a.0.3
b.0.24
c.2.2
d.0.44
e.0.2
f.0.68
g.0.78

Câu 41:

Giải thuật Kmeans được thực hiện theo các bước sau:

1. Khởi động ngẫu nhiên K tâm (center) của K clusters
2. Mỗi phần tử được gán cho tâm gần nhất với phần tử dựa vào khoảng cách (e.g. khoảng cách Euclid)
3. Cập nhật lại các tâm của K clusters, mỗi tâm là giá trị trung bình (mean) của các phần tử trong cluster của nó
4. Lặp lại bước 2,3 cho đến khi hội tụ

Anh/chị hãy cho biết ý nào sau đây là phát biểu đúng?

a. Giải thuật Kmeans cần biết trước số lượng nhóm cần phải gom cụm
b. Các phần tử cùng nhóm có tính chất khác nhau
c. Sử dụng khoảng cách Euclid để tìm các phần tử tương tự nhau cho các thuộc tính có kiểu liệt kê.
d. Hội tụ nghĩa là các phần tử gom về cùng 1 nhóm

Câu 42:

Đại học ABC từ năm 1976 đến nay đã sưu tập và lưu giữ hầu hết các giống lúa mùa cổ truyền của vùng ĐBSCL. Số lượng cụ thể là hơn 1.900 mẫu. Mỗi một giống lúa trong ngân hàng thông tin được mô tả bởi 69 đặc điểm hình thái, đặc tính nông sinh học (ví dụ: góc lá đòng, độ cứng cây, chiều cao cây, thời gian sinh trưởng, số bông hữu hiệu/khóm, số hạt chắc, lép/bông, khối lượng nghìn hạt, năng suất lý thuyết - theo thang điểm của International Rice Research Institute). Chúng ta cần tạo ra các công cụ tin học giúp cho các nhà nghiên cứu có thể phát hiện ra các mẫu lúa cùng nhóm với nhau. Giải thuật máy học nào sau đây phù hợp cho bài toán này?

a. Giải thuật học có giám sát, giải thuật hồi quy (regression).
b. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm (clustering).
c. Giải thuật học có giám sát, giải thuật phân lớp (classification).
d. Giải thuật học không có giám sát, giải thuật hồi quy (regression).

Câu 43:

Hiện tượng overfitting (học vẹt) xảy ra trong tình huống nào sau đây?

a. Xảy ra khi sai số dự đoán của mô hình trên cả tập huấn luyện và tập kiểm thử đều thấp
b. Xảy ra khi sai số dự đoán của mô hình trên cả tập huấn luyện và tập kiểm thử đều cao.
c. Xảy ra khi sai số dự đoán của mô hình trên tập huấn luyện thấp, nhưng trên tập kiểm thử thì cao
d. Xảy ra khi sai số dự đoán của mô hình trên tập huấn luyện cao, nhưng trên tập kiểm thử thì thấp

Câu 44:

Phát biểu nào sau đây là đúng trong việc chọn lựa các mô hình trong phương pháp tập hợp mô hình?

1. Các mô hình có thể sử dụng cùng một giải thuật với các tham số khác nhau.
2. Các mô hình có thể được xây dựng từ các giải thuật khác nhau
3. Các mô hình có thể được xây dựng trên những tập dữ liệu huấn luyện khác nhau.

a. Câu 2 đúng
b. Câu 1 đúng
c. Câu 1 và 3 đúng
d. Cả 1, 2, 3 đều đúng

Câu 45:

Cho tập dữ liệu như hình bên dưới. Người ta mong muốn xây dựng một phần mềm có khả năng dự đoán lượng khí thải CO2 từ một chiếc ô tô khi chúng ta chỉ biết trọng lượng (Weight) và thể tích (Volume) của nó.

Car	Model	Volume	Weight	CO2
Toyota	Aygo	1.0	790	99
Mitsubishi	Space Star	1.2	1160	95
Skoda	Citigo	1.0	929	95
Fiat	500	0.9	865	90
Mini	Cooper	1.5	1140	105
VW	Up!	1.0	929	105
Skoda	Fabia	1.4	1109	90
Mercedes	A-Class	1.5	1365	92
Ford	Fiesta	1.5	1112	98
Audi	A1	1.6	1150	99

Người ta sử dụng 4 giải thuật hồi quy khác nhau để xây dựng mô hình dự báo. Để đánh giá hiệu quả của giải thuật, người ta đo chỉ số MAE (Mean absolute error) và có kết quả như sau:

STT	Giải thuật	MAE
1	Giải thuật 1	10.4
2	Giải thuật 2	20.7
3	Giải thuật 3	9.8
4	Giải thuật 4	15.6

Với các giá trị MAE như trên, giải thuật nào là phù hợp nhất cho bài toán này?

a. Giải thuật 4
b. Giải thuật 2
c. Giải thuật 3
d. Giải thuật 1

Câu 46:

Cho tập dữ liệu sau:

	Màu tóc	Màu mắt	Chiều cao	Cân nặng	Trình độ
Nam	Đen	Đen	Cao	Trung bình	Cao đẳng
Mai	Nâu	Đen	Thấp	Trung bình	Đại học
Lan	Nâu	Đen	Cao	Trung bình	Cao đẳng
Cúc	Bạch kim	Nâu	Thấp	Nhẹ cân	Thạc sĩ
Trúc	Đen	Đen	Thấp	Nhẹ cân	Cao đẳng

Hãy tính khoảng cách giữa Nam đối với 4 bạn nữ Mai, Lan, Cúc, Trúc và cho biết Nam có khoảng cách "gần" với bạn nữ nào nhất?

a. Trúc
b. Lan
c. Mai
d. Cúc

Câu 47:

Cho tập dữ liệu như bảng bên dưới gồm 6 phần tử với 2 thuộc tính màu tóc, độ bóng mượt và nhãn là tình trạng tóc. Anh/chị hãy xác định chỉ số Gini của thuộc tính "màu tóc"

a.0.25
b.1
c.0.5
d.0.44

STT	Màu tóc	Độ bóng mượt	Tình trạng tóc
1	Bạch kim	49	Bị cháy nắng
2	Đen	55	Bị cháy nắng
3	Bạch kim	67	Bị cháy nắng
4	Bạch kim	65	Không cháy nắng
5	Bạch kim	62	Không cháy nắng
6	Đen	70	Không cháy nắng

Câu 48:

Cho tập dữ liệu như bảng bên dưới gồm 7 phần tử, mỗi phần tử có 2 thuộc tính (Bao gồm tiền điện, kích thước phòng) và cột nhãn/lớp là cột "giá thuê phòng".

STT	Bao gồm tiền điện	Kích thước phòng	Giá thuê phòng
1.	Không	10	Thấp
2.	Không	17	Cao
3.	Không	14	Trung bình
4.	Có	17.5	Cao
5.	Có	13	Trung bình
6.	Có	16.5	Cao
7.	Có	8	Thấp

Với mô hình Bayes thơ ngây được xây dựng sẵn như bảng bên dưới, anh/chị hãy cho biết giá trị u của bảng "bao gồm tiền điện" ở cột [Không, Cao] là bao nhiêu?

Bao gồm tiền điện	Thấp	Trung bình	Cao
Không	x	t	**u**
Có	v	z	w

a.0.33
b.0.67
c.0.5
d.Tất cả đều sai

Câu 49:

Để đánh giá hiệu quả của bài toán phân lớp nhị phân với số lượng lớp quan tâm có tỉ lệ rất ít so với phân lớp còn lại, ta sử dụng chỉ số nào sau đây?
a. Precision
b. MSE hoặc RMSE hoặc MAE
c. Accuracy
d. Accuracy và F1

Câu 50:

Phát biểu nào sau đây về Naive Bayes là không đúng?

a. Các thuộc tính có thể là chữ hoặc số.
b. Các thuộc tính đều quan trọng như nhau.
c. Các thuộc tính phụ thuộc thống kê lẫn nhau với giá trị của lớp.
d. Các thuộc tính độc lập về mặt thống kê với nhau theo giá trị của lớp.

Câu 51:

Dựa vào thông tin tuổi, giới tính, nồng độ cholesterol, nhịp tim, điện tâm đồ người ta cần dự báo mức độ bệnh của bệnh nhân với các cấp độ sau: bình thường, nhẹ, nặng và nghiêm trọng. Theo anh/chị, chúng ta nên sử dụng giải thuật gì để dự báo mức độ bệnh của bệnh nhân?

a. Giải thuật học không có giám sát, giải thuật phân lớp.
b. Giải thuật học có giám sát, giải thuật hồi quy.
c. Giải thuật học có giám sát, giải thuật phân lớp.
d. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm.

Câu 52:

Cho 03 câu A, B, C như sau:

- A = "We are learning Natural Language Processing"
- B = "We are learning Data Science"
- C = "Natural Language Processing comes under Data Science"

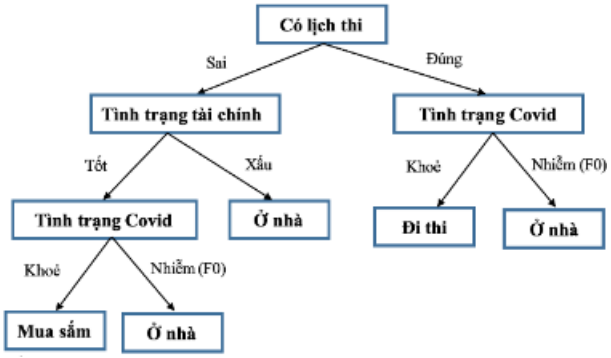
Hãy sử dụng Bag of Word để tính vector của A:
Biết từ điển gồm:

- [are, comes, data, language, learning, natural, processing, science, under, we]

a. 1,0,1,0,1,0,0,1,0,1
b. 1,0,1,0,1,1,1,0,0,1
c. 1,0,0,1,1,1,1,0,0,1
d. 0,1,1,1,0,1,1,1,1,0

Câu 53:

Cho cây quyết định như hình bên dưới, NHỮNG phát biểu sau đây là SAI:



a. Nhãn của tập dữ liệu: có lịch thi, tình trạng tài chính, tình trạng Covid
b. Nhãn của tập dữ liệu: mua sắm, ở nhà, đi thi
c. Nhãn của tập dữ liệu: mua sắm, ở nhà, đi thi, đúng, sai, tốt, xấu, nhiễm (F0), Khoẻ
d. Nhãn của tập dữ liệu: đúng, sai, tốt, xấu, nhiễm (F0), Khoẻ

1.(a), (b) và (d)
2.(a), (b) và (c)
3.(b), (c) và (d)
4.(a), (c) và (d)

Câu 54:

Cho dữ liệu như bảng sau

A1	A2	A3
9	4	6.5
5	3	5
8	2	1.5
7	4	5.5
6	2	5

Người ta thực hiện gom nhóm dữ liệu trên thành 2 nhóm bằng giải thuật Kmeans, với các thông tin sau:

- Các tâm khởi động ngẫu nhiên : c1(6,3,5) ; c2(5,2,1)
- Khoảng cách sử dụng: khoảng cách Manhatan

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

Anh/ chị hãy cho biết khoảng cách từ phần tử thứ 3 (Tô màu) đến tâm C1 và C2 là bao nhiêu và xác định nhóm của phần tử thứ 3 này là nhóm 1 hay nhóm 2?

a. Khoảng cách đến tâm C1= 4.15, đến tâm C2=4.36; thuộc nhóm 1
b. Khoảng cách đến tâm C1= 6.5, đến tâm C2=3.5; thuộc nhóm 1
c. Khoảng cách đến tâm C1= 4.15, đến tâm C2=4.36; thuộc nhóm 2
d. Khoảng cách đến tâm C1= 6.5, đến tâm C2=3.5; thuộc nhóm 2

Câu 55:

Các phát biểu nào sau đây là phát biểu đúng?

a. Giải thuật KNN không có quá trình học nên khi phân loại có tốc độ nhanh hơn các giải thuật khác
b. Giải thuật KNN giải quyết cho bài toán phân loại và bài toán gom nhóm
c. Giải thuật KNN không có quá trình học nên khi phân loại có tốc độ chậm hơn các giải thuật khác

Câu 56:

Cho tập dữ liệu như sau:

Sử dụng giải thuật KNN với K=5, anh chị hãy dự đoán cân nặng của anh Nguyễn Văn A có chiều cao 1.71 m?

a.61.29
b.64.65
c.64.47
d.66.28
e.63.11
f.68.1

Height (m)	Weight (kg)
1.47	52.21
1.5	53.12
1.52	54.48
1.55	55.84
1.57	57.2
1.6	58.57
1.63	59.93
1.65	61.29
1.68	63.11
1.7	64.47
1.73	66.28
1.75	68.1
1.78	69.92
1.8	72.19
1.83	74.46

Câu 57:

Phương pháp tập hợp mô hình (Ensemble Methods) là kết hợp nhiều mô hình cơ sở dựa trên tập học nhằm cải thiện độ chính xác của giải thuật dự đoán. Tăng hiệu quả của mô hình dựa trên cơ sở giảm lỗi bias, variance. Lỗi variance được hiểu như thế nào?

a. Lỗi variance là thành phần lỗi do biến động liên quan đến sự ngẫu nhiên của tập học
b. Lỗi variance là lỗi liên quan đến mô hình (bộ phân lớp/ dự đoán) mà không liên quan đến dữ liệu được dùng để huấn luyện
c. Lỗi variance là lỗi liên quan đến việc chọn lựa mô hình không phù hợp
d. Tất cả đều đúng

Câu 58:

Cho dữ liệu như bảng sau (dữ liệu không bao gồm cột STT)

STT	A1	A2	A3
1	7.9	4	6.4
2	5	2.7	5.1
3	4.5	2	1.5
4	7	3.5	5.5
5	6	3	5

Người ta thực hiện gom nhóm dữ liệu trên thành 2 nhóm bằng giải thuật Kmeans, với các thông tin sau:

- Các tâm khởi động ngẫu nhiên: c1(6, 3, 5); c2(4.5, 2, 1.5)
- Khoảng cách sử dụng: khoảng cách Manhatan

Người ta đã tính được khoảng cách tới các tâm như bảng bên dưới, anh/ chị hãy cho biết tâm C1 của bước tiếp theo có giá trị là bao nhiêu?

STT	Khoảng cách đến tâm C1	Khoảng cách đến tâm C2
1	4.3	10.3
2	1.4	4.8
3	6	0
4	2	8
5	0	6

a. Tâm C1(3.25; 5.375)
b. Tâm C1(6.475; 3.3, 5.5)
c. Tâm C1(2; 7.25)
d. Tâm C1(6.625; 3.25)

Câu 59:

Giả sử người ta có n biến độc lập (X1, X2, …, Xn) và một biến phụ thuộc Y. Nếu sử dụng mô hình hồi quy thì ta cần xác định những thông tin nào sao đây?

a. Hai tham số theta0 và theta1
b. Các tham số theta1, theta2, …, thetan
c. Các tham số theta0, theta1, theta2, …, thetan
d. Ba tham số theta1, theta2, theta3

Câu 60:

Cho một tập dữ liệu chứa các thông tin sau về rượu vang: Độ Ph, độ cồn, lượng đường, axit cố định, axit bay hơi, thời gian xuất xưởng. Người ta cần xác định thời gian xuất xưởng của từng loại rượu theo đơn vị giờ. Theo anh/chị các thuộc tính của tập dữ liệu này là gì?

a. Thông tin thuộc tính: thời gian xuất xưởng của rượu.
b. Thông tin thuộc tính: Độ Ph, độ cồn, hàm lượng đường, axit cố định, axit bay hơi của rượu.
c. Thông tin thuộc tính: Độ Ph, độ cồn, hàm lượng đường, thời gian xuất xưởng của rượu.
d. Thông tin thuộc tính: Độ Ph, độ cồn, hàm lượng đường, axit cố định, axit bay hơi của rượu, thời gian xuất xưởng của rượu.

Câu 61:

Máy học là một nhánh của lĩnh vực học nào sau đây?

a. Học có giám sát
b. Học tăng cường
c. Trí tuệ nhân tạo
d. Học sâu

Câu 62:

Phát biểu nào sau đây là đúng đối với giải thuật K láng giềng gần (KNN)?

a. Nên chọn giá trị K lẻ để việc phân lớp được hiệu quả điểm
b. Thời gian thực hiện giải thuật KNN nhanh hơn các giải thuật khác do không cần xây dựng mô hình
c. Độ chính xác của việc phân lớp hiệu quả hơn khi chọn giá trị K nhỏ
d. Giải thuật KNN không phù hợp cho bài toán hồi quy

Câu 63:

Theo anh/chị các định nghĩa nào sau đây là phát biểu đúng

a. Máy học là chương trình máy tính cho phép học tự động từ dữ liệu để nhận dạng các mẫu phức tạp, tạo ra hành vi ứng xử thông minh với trường hợp mới đến
b. Học không giám sát là thuật toán học thực hiện mô hình hoá một tập dữ liệu đầu vào, không được gán nhãn (lớp, giá trị cần dự báo)
c. Học có giám sát là thuật toán học tạo ra một hàm ánh xạ dữ liệu đầu vào tới kết quả đích mong muốn (nhãn, lớp, giá trị cần dự báo). Trong học có giám sát, tập dữ liệu dùng để huấn luyện phải được gán nhãn, lớp hay giá trị cần dự báo
d. Tất cả đều đúng

Câu 64:

Công ty viễn thông ABC dự định phát triển API phát hiện các tin nhắn rác. Công ty đã tiến hành thu thập 10.000 mẫu dữ liệu, bao gồm 3.000 tin nhắn rác và 7.000 tin nhắn bình thường. Chúng ta cần sử dụng giải thuật máy học gì để giúp công ty ABC xây dựng mô hình dự đoán tin nhắn rác.

a. Giải thuật học có giám sát, giải thuật hồi quy (regression).
b. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm (clustering).
c. Tìm kiếm đối kháng
d. Giải thuật học có giám sát, giải thuật phân lớp (classification).

Câu 65:

Anh/chị hãy lựa chọn giải thuật có thể sử dụng cho cả bài toán hồi quy và bài toán phân lớp.

a. Rừng ngẫu nhiên
b. Hồi quy tuyến tính
c. Kmeans
d. Bayes thơ ngây

Câu 66:

Tại bước huấn luyện mô hình của giải thuật Bayes thơ ngây người ta cần làm gì?

a. Tính xác suất xuất hiện của các thuộc tính theo từng giá trị nhẫn
b. Tính xác suất xuất hiện của các thuộc tính theo từng giá trị nhẫn và xác suất xuất hiện của từng nhẫn
c. Tính xác suất xuất hiện của nhãn
d. Tất cả đều sai

Câu 67:

Anh/chị hãy tính giá trị F1, Accuracy của kết quả đánh giá mô hình sau
Ma trận confusion từ mô Hình M2

Dự báo =>	Dương	Âm
Dương	2	7
Âm	30	29961

a. Accuracy = 90%, F1 = 20%
b. Accuracy = 99.9%, F1= 9.8%
c. Accuracy = 90%, F1=9.8%
d. Accuracy = 99.9%, F1=20%

Câu 68:

Giả sử tập dữ liệu có 30000 mẫu tin trong đó có 15 mẫu tin thuộc lớp dương (+1), có hai mô hình phân lớp M1 và M2 cho kết quả tương ứng trong bảng 1, 2 như bên dưới. Anh/chị hãy cho biết mô hình nào thích hợp để xử lý tập dữ liệu trên? Hãy giải thích lý do cho lựa chọn đó.

Ma trận confusion từ mô Hình M1		
Dự báo =>	Dương	Âm
Dương	13	2
Âm	92	29893

Ma trận confusion từ mô Hình M2		
Dự báo =>	Dương	Âm
Dương	3	12
Âm	50	29935

a. Chọn mô hình M2 vì recall của M2 > M1
b. Chọn mô hình M2 vì accuracy của M2 > M1
c. Chọn mô hình M1 vì F1 của M1 > M2 => Dữ liệu mất cân bằng
d. Chọn mô hình M1 vì precision của M1 > M2

Câu 69:

Cho ma trận confusion sau, anh chị hãy cho biết độ chính xác của rượu loại 3

Phân loại rượu	Loại 1	Loại 2	Loại 3	Loại 4
Loại 1	54	5	1	0
Loại 2	0	67	3	0
Loại 3	0	5	85	10
Loại 4	0	0	8	92

a. 90%
b. 88%
c. 70%
d. 85%

Câu 70:

Cho tập dữ liệu như bằng bên dưới

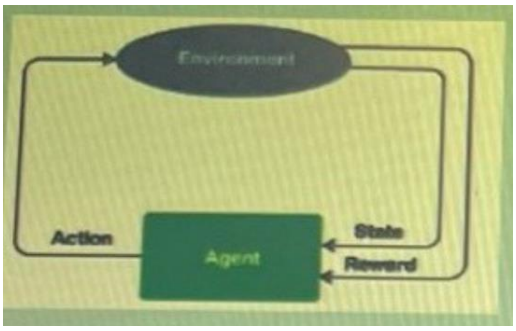
STT	X_1	X_2	Y
1	3	9	9
2	5	8	10
3	8	10	15

Giả sử sau n bước lập, ta tìm được giá trị theta0 = 0.3; theta1 = 0.46; theta2 = 0.32, anh/chị hãy cho biết giá trị dự báo giá trị Y nếu biết X1 = 5, X2 = 12

a 17
b. 6.14
c 7.02
d. 6.44

Câu 71:

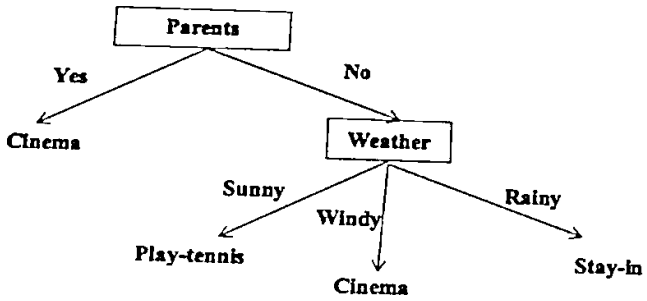
Theo anh/chị sơ đồ bên dưới mô tả cho phương pháp nào?



a. Học tăng cường
b. Học bán giám sát
c. Học có giám sát
d. Học không có giám sát

Câu 72:

Cho cây quyết định như hình bên dưới



Hãy xác định nhãn cho trường hợp sau: Weather = Rainy, Parents = Yes

a. Play-tennis
b. Cinema
c. Stay-in
d. Tất cả đều đúng

Câu 73:

Cho tập dữ liệu như bảng bên dưới gồm 7 phần tử, mỗi phần tử có 2 thuộc tính (Bao gồm tiền điện, kích thước phòng) và cột nhân/lớp là cột "giá thuê phòng"

STT	Bao gồm tiền điện	Kích thước phòng	Giá thuê phòng
1.	Không	10	Thấp
2.	Không	17	Cao
3.	Không	14	Trung bình
4.	Có	17.5	Cao
5.	Có	13	Trung bình
6.	Có	16.5	Cao
7.	Có	8	Thấp

Với mô hình Bayes thơ ngây được xây dựng sẵn như bảng bên dưới, anh/chị hãy cho biết giá trị x của bảng "kích thước phòng" ở ô [Phương sai, Thấp] là bao nhiêu?

Kích thước phòng	Thấp	Trung bình	Cao
mean	t	u	v
Phương sai (σ^2)	x	y	z

a. 0.5
b. 0.25
c. 2
d. 0.71

Câu 74:

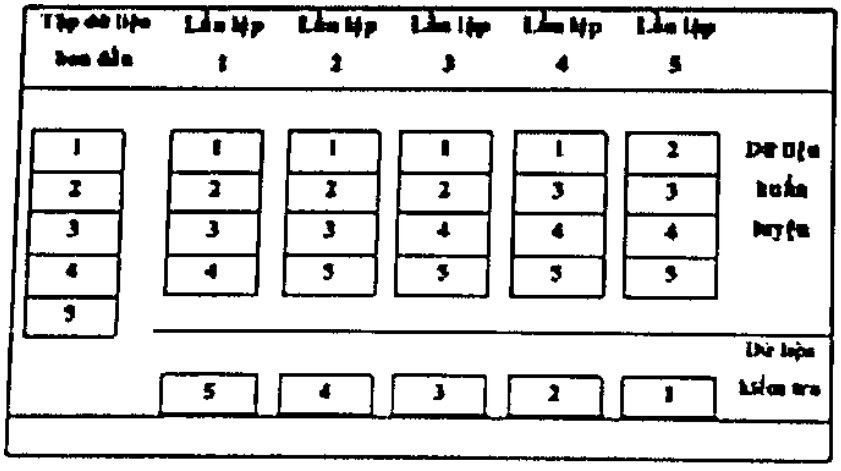
Cho bảng tính sẵn xác suất cho nhãn "PP1" và "PP2" của thuộc tính "tình trạng bệnh" như bảng bên dưới, các giá trị xác suất nào bên dưới là chính xác nếu sử dụng ước lượng Laplace để tránh tình trạng tạo ra giá trị 0 ở ô [mức 1, PP1]

Tình trạng bệnh	PP1	PP2
Mức 1	0/4	2/3
Mức 2	2/4	1/3
Mức 3	2/4	0/3

a. Mức 1 = (1/3 + 0) / (4+1); Mức 2 = (1/3 + 2) / (4+ 1); Mức 3= (1/3 + 2)/(4+1)
b. Mức 1 = (1/5 + 0) / (4+1); Mức 2 = (2/5+2) / (4+1); Mức 3= (2/5+2)/(4+1)
c. Mức 1 = (1/6 + 0) / (4+1); Mức 2 = (2/6 + 2) / (4+1); Mức 3= (3/6 + 2) / (4+1)
d. Tất cả đều đúng

Câu 75:

Anh/chị hãy xác định nghi thức đánh giá cho hình minh hoạ bên dưới:



a. Tất cả đều sai
b. K-fold với k=5
c. Hold-out với tập huấn luyện chiếm 4/5 và tập kiểm tra 1/5