CHƯƠNG 1: CÂU HỎI LÝ THUYẾT

Câu 1) Trình bày thuộc tính rời rạc và thuộc tính liên tục. Lấy ví dụ để chứng tỏ sự khác biệt.

Thuộc tính rời rạc

- Chỉ có một tập giá trị hữu hạn hoặc vô hạn đếm được
  - Ví dụ 1: Mã zip, nghề nghiệp hoặc tập hợp từ trong tập hợp tài liệu
  - Ví dụ 2: Số học sinh của một lớp không thể có nửa học sinh
  - Ví dụ 3: Kết quả tung 2 viên xúc xắc Chỉ có các giá trị 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 ta không thể có 2,1 hoặc 3,5
- Đôi khi, được biểu diễn dưới dạng biến số nguyên
- Thuộc tính nhị phân là trường hợp đặc biệt của thuộc tính rời rạc

Thuộc tính liên tục

- Có giá trị thuộc tính là số thực
  - Ví dụ: Nhiệt độ, chiều cao, cân nặng
- Trên thực tế, giá trị thực chỉ có thể được đo và biểu diễn bằng một số hữu hạn chữ số
- Thuộc tính liên tục thường được biểu diễn dưới dạng biến dấu phẩy động

Câu 2) Trình bày ưu, nhược điểm của Cây quyết định.

Thuận lợi:

- Nó có thể được sử dụng cho cả các vấn đề phân loại và hồi quy: Cây quyết định có thể được sử dụng để dự đoán cả giá trị liên tục và rời rạc, tức là chúng hoạt động tốt trong cả hồi quy và nhiệm vụ phân loại.
- Vì cây quyết định rất đơn giản nên chúng đòi hỏi ít nỗ lực hơn để hiểu thuật toán.
  - Nó có thể nắm bắt các mối quan hệ phi tuyến tính: Chúng có thể được sử dụng để phân loại các mối quan hệ phi tuyến tính có thể phân tách dữ liệu.
- Ưu điểm của thuật toán cây quyết định là nó không yêu cầu bất kỳ phép biến đổi nào của các tính năng nếu chúng ta đang xử lý dữ liệu phi tuyến tính vì cây quyết định không có nhiều tính đến các kết hợp có trọng số cùng một lúc.
- Chúng rất nhanh và hiệu quả so với KNN và các thuật toán phân loại khác.
- Dễ hiểu, dễ hiểu, dễ hình dung.

Kiểu dữ liệu của cây quyết định có thể xử lý bất kỳ loại dữ liệu nào dù là số hay phân loại hoặc boolean.

Việc chuẩn hóa không cần thiết trong Cây quyết định.

Cây quyết định là một trong những thuật toán học máy mà chúng ta không phải lo lắng về mở rộng tính năng. Một cái khác là rừng ngẫu nhiên. Những thuật toán đó là bất biến quy mô. Nó cho chúng ta một ý tưởng hay về tầm quan trọng tương đối của các thuộc tính.

Hữu ích trong việc khám phá dữ liệu: Cây quyết định là một trong những cách nhanh nhất để xác định các biến có ý nghĩa và mối quan hệ giữa hai hoặc nhiều biến. Cây quyết định tốt hơn sức mạnh mà chúng ta có thể tạo các biến/tính năng mới cho biến kết quả.

Cần ít chuẩn bị dữ liệu hơn: Trong cây quyết định, không có tác động nào từ bên ngoài hoặc thiếu dữ liệu ở nút của cây, đó là lý do tại sao cây quyết định cần ít dữ liệu hơn.

Cây quyết định không tham số: Phương pháp phi tham số được định nghĩa là phương pháp trong đó không có giả định nào về sự phân bố không gian và cấu trúc phân loại.

Nhược điểm:

Về việc phân chia cây quyết định cho biến số hàng triệu bản ghi: Thời gian độ phức tạp để vận hành hoạt động này là rất lớn và tiếp tục tăng lên khi số lượng số bản ghi được tăng lên cây quyết định với các biến số mất rất nhiều thời gian cho đào tạo.

Tương tự, điều này xảy ra trong các kỹ thuật như rừng ngẫu nhiên, XGBoost.

Cây quyết định cho nhiều tính năng: Mất nhiều thời gian hơn để độ phức tạp của thời gian huấn luyện tăng lên khi đầu vào tăng lên.

Trồng cây từ tập huấn luyện: Cắt tỉa quá mức (trước, sau), phương pháp tổng hợp rừng ngẫu nhiên

Phương pháp overfitting: Nếu bàn về overfitting thì đây là một trong những phương pháp khó nhất đối với mô hình cây quyết định Vấn đề trang bị quá mức có thể được giải quyết bằng cách thiết lập các ràng buộc trên mô hình tham số và phương pháp cắt tỉa.

Như bạn đã biết, cây quyết định thường cần quá nhiều dữ liệu. Trong vấn đề trang bị quá mức, có sự chênh lệch rất cao về kết quả đầu ra dẫn đến nhiều sai sót trong ước tính cuối cùng và có thể cho thấy kết quả đầu ra có độ thiếu chính xác cao. Đạt được độ lệch bằng 0 (trang bị quá mức), dẫn đến phương sai cao.

Khả năng tái sử dụng trong cây quyết định: Trong cây quyết định có những biến thể nhỏ trong dữ liệu có thể xuất ra một cây phức tạp khác được tạo ra. Điều này được gọi là sự khác biệt trong cây quyết định, có thể được giảm bớt bằng một số phương pháp như đóng bao và tăng tốc.

Nó không thể được sử dụng trong dữ liệu lớn: Nếu kích thước của dữ liệu quá lớn thì một cây có thể phát triển rất nhiều của các nút có thể dẫn đến sự phức tạp và dẫn đến trang bị quá mức.

Không có gì đảm bảo sẽ trả lại cây quyết định hiệu quả 100%.

Câu 3) Trình bày ưu, nhược điểm của Máy Vector Hỗ trợ.

Thuận lợi:

Cho kết quả tốt ngay cả khi không có đủ thông tin về dữ liệu. Cũng hoạt động tốt với dữ liệu phi cấu trúc.

Giải quyết các vấn đề phức tạp bằng chức năng giải kernel tiện lợi.

Khả năng mở rộng tương đối tốt dữ liệu nhiều chiều.

Bộ phân loại SVM hoạt động tốt trong không gian nhiều chiều và có độ chính xác tuyệt vời. SVM bộ phân loại yêu cầu ít bộ nhớ hơn vì chúng chỉ sử dụng một phần dữ liệu huấn luyện.

SVM hoạt động khá tốt khi có khoảng cách lớn giữa các lớp.

Không gian nhiều chiều phù hợp hơn với SVM.

Khi số lượng kích thước vượt quá số lượng mẫu, SVM rất hữu ích.

SVM sử dụng bộ nhớ hiệu quả.

Nhược điểm:

Rất khó để chọn chức năng giải pháp kernel phù hợp.

Thời gian huấn luyện dài khi sử dụng tập dữ liệu lớn.

Có thể khó diễn giải và hiểu được vì các vấn đề do yếu tố cá nhân gây ra

và trọng số của các biến.

Trọng số của các biến số không phải là hằng số, do đó sự đóng góp của mỗi biến vào đầu ra là biến thể.

SVM cần thời gian đào tạo dài; kết quả là nó không thực tế đối với các tập dữ liệu lớn.

Việc các bộ phân loại SVM không có khả năng xử lý các lớp chồng chéo là một nhược điểm khác.

Bộ dữ liệu lớn không phù hợp với thuật toán SVM.

Khi tập dữ liệu chứa nhiều nhiễu hơn, chẳng hạn như các lớp mục tiêu chồng chéo, SVM không biểu diễn là tốt.

SVM sẽ hoạt động kém khi số lượng đặc trưng cho mỗi điểm dữ liệu lớn hơn số lượng mẫu dữ liệu huấn luyện

Câu 4) Cho một ví dụ về ứng dụng khai thác dữ liệu trong lĩnh vực giáo dục hoặc công cộng vận tải. Dựa trên ví dụ của bạn, loại dữ liệu và phương pháp khai thác dữ liệu nào bạn có thể sử dụng?

Câu 5) So sánh học có giám sát và học không giám sát. Lấy ví dụ để chứng tỏ sự khác biệt.

Học có giám sát: là một phương pháp học máy được xác định bằng cách sử dụng các bộ dữ liệu được gắn nhãn. Những cái này các bộ dữ liệu được thiết kế để huấn luyện hoặc “giám sát” các thuật toán trong việc phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác. Bằng cách sử dụng đầu vào và đầu ra được gắn nhãn, mô hình có thể đo lường độ chính xác và học hỏi theo thời gian.

Học có giám sát có thể được chia thành hai loại vấn đề khi khai thác dữ liệu: phân loại và hồi quy:

Các bài toán phân loại sử dụng thuật toán để gán chính xác dữ liệu kiểm tra vào các phần cụ thể phân loại, chẳng hạn như tách táo khỏi cam.

Hồi quy là một loại phương pháp học có giám sát khác sử dụng thuật toán để hiểu mối quan hệ giữa các biến phụ thuộc và biến độc lập. Không được giám sát học tập: sử dụng thuật toán học máy để phân tích và phân cụm các tập dữ liệu chưa được gắn nhãn. Những cái này thuật toán khám phá các mẫu ẩn trong dữ liệu mà không cần sự can thiệp của con người

Học không giám sát: sử dụng thuật toán học máy để phân tích và phân cụm các tập dữ liệu không được gắn nhãn.
Các thuật toán này khám phá các mẫu ẩn trong dữ liệu mà không cần sự can thiệp của con người

Các mô hình học không giám sát được sử dụng cho ba nhiệm vụ chính: phân cụm, liên kết và tính chiều
sự giảm bớt:

- Phân cụm là một kỹ thuật khai thác dữ liệu để nhóm các dữ liệu chưa được gắn nhãn dựa trên sự giống nhau của chúng hoặc sự khác biệt.
- Liên kết là một loại phương pháp học không giám sát khác sử dụng các quy tắc khác nhau để tìm mối quan hệ giữa các biến trong một tập dữ liệu nhất định.
- Giảm kích thước là một kỹ thuật học được sử dụng khi số lượng các tính năng (hoặc kích thước) trong một tập dữ liệu nhất định quá cao

Sự khác biệt chính giữa hai phương pháp này là việc sử dụng các tập dữ liệu được dán nhãn. Nói một cách đơn giản,
học có giám sát sử dụng dữ liệu đầu vào và đầu ra được gắn nhãn, trong khi thuật toán học không giám sát thì
không.

Trong học có giám sát, thuật toán "học" từ tập dữ liệu huấn luyện bằng cách thực hiện lặp đi lặp lại
dự đoán về dữ liệu và điều chỉnh để có câu trả lời đúng.

- Ví dụ: mô hình học tập có giám sát có thể dự đoán thời gian đi làm của bạn vào thời điểm trong ngày, điều kiện thời tiết, v.v. Nhưng trước tiên, bạn sẽ phải huấn luyện nó để biết rằng trời mưa kéo dài thời gian lái xe.

Ngược lại, các mô hình học không giám sát hoạt động độc lập để khám phá cấu trúc vốn có của
dữ liệu không có nhãn.

- Ví dụ: mô hình học tập không giám sát có thể xác định rằng người mua hàng trực tuyến thường mua các nhóm sản phẩm cùng một lúc. Tuy nhiên, một nhà phân tích dữ liệu sẽ cần phải xác nhận rằng công cụ đề xuất có thể nhóm quần áo trẻ em với một thứ tự tã lót, nước sốt táo và cốc tập uống.

Câu 6) Cho một ví dụ về ứng dụng khai thác dữ liệu trong lĩnh vực y tế hoặc kinh tế. Dựa trên
trong ví dụ của bạn, bạn có thể sử dụng loại dữ liệu và phương pháp khai thác dữ liệu nào?

Câu 7) Trình bày khái niệm về phân cụm và phân loại. Lấy ví dụ để thể hiện ý tưởng của bạn.

Phân cụm

Phân cụm là phương pháp chuyển đổi một nhóm đối tượng trừu tượng thành các lớp có tính chất tương tự các đối tượng.

Phân cụm là phương pháp phân chia một tập hợp dữ liệu hoặc đối tượng thành một tập hợp quan trọng các lớp con được gọi là cụm.

Nó giúp người dùng hiểu cấu trúc hoặc nhóm tự nhiên trong một tập dữ liệu và được sử dụng như một công cụ độc lập để hiểu rõ hơn về phân phối dữ liệu hoặc như một bước tiền xử lý cho các thuật toán khác

Phân loại

Phân loại là chức năng khai thác dữ liệu gán các mục trong bộ sưu tập cho các danh mục mục tiêu hoặc các lớp học.

Mục tiêu của việc phân loại là dự đoán chính xác lớp mục tiêu cho từng trường hợp trong dữ liệu.

Phân loại dữ liệu (xây dựng mô hình) dựa trên tập huấn luyện và các giá trị (nhãn lớp) trong một thuộc tính phân loại và sử dụng nó trong việc phân loại dữ liệu mới.

Câu 8) Lấy một cặp thuộc tính mà chúng có mối tương quan cao.

(Mạnh, Yếu) (Dài, Ngắn) (Lớn, Nhỏ) (Sớm, Muộn)

Câu 9) Bạn có nghĩ “Tương quan có hàm ý quan hệ nhân quả” không? và giải thích tại sao.

Tuy nhiên, mối tương quan giữa các biến không tự động có nghĩa là sự thay đổi của một biến là nguyên nhân làm thay đổi giá trị của biến kia. Nguyên nhân chỉ ra rằng một sự kiện là kết quả của việc xảy ra sự kiện kia; tức là có mối quan hệ nhân quả giữa hai sự kiện.

Câu 10) Dữ liệu không đầy đủ (thiếu) là gì?

Dữ liệu không phải lúc nào cũng có sẵn

Ví dụ: Nhiều bộ không có giá trị được ghi lại cho một số thuộc tính, chẳng hạn như thu nhập của khách hàng khi bán hàng dữ liệu

Thiếu dữ liệu có thể do:

trục trặc thiết bị

không phù hợp với dữ liệu được ghi khác và do đó bị xóa

dữ liệu không được nhập do hiểu nhầm

một số dữ liệu nhất định có thể không được coi là quan trọng tại thời điểm nhập cảnh

không đăng ký lịch sử hoặc thay đổi dữ liệu

Dữ liệu bị thiếu có thể cần được suy luận

Câu 11) Làm thế nào để xử lý dữ liệu không đầy đủ (thiếu)?

- Bỏ qua bộ dữ liệu: thường được thực hiện khi thiếu nhãn lớp (khi thực hiện phân loại)—không hiệu quả khi % giá trị còn thiếu trên mỗi thuộc tính thay đổi đáng kể
- Điền thủ công giá trị còn thiếu: tẻ nhạt + không khả thi?
- Tự động điền vào đó:
  - hằng số toàn cục
  - thuộc tính trung bình
  - giá trị trung bình của thuộc tính đối với tất cả các mẫu thuộc cùng một lớp: thông minh hơn
  - giá trị có khả năng xảy ra cao nhất: dựa trên suy luận như công thức Bayes hoặc cây quyết định

Câu 12) Dữ liệu nhiễu là gì?

- Nhiễu: sai số hoặc phương sai ngẫu nhiên của một biến đo được
- Giá trị thuộc tính không chính xác có thể là do:
  - công cụ thu thập dữ liệu bị lỗi
  - vấn đề nhập dữ liệu
  - vấn đề truyền dữ liệu
  - hạn chế về công nghệ
  - mâu thuẫn trong quy ước đặt tên
- Các sự cố dữ liệu khác yêu cầu làm sạch dữ liệu
  - bản ghi trùng lặp
  - dữ liệu không đầy đủ
  - dữ liệu không nhất quán

Câu 13) Làm thế nào để xử lý dữ liệu nhiễu?

- Thùng rác
  - Đầu tiên sắp xếp dữ liệu và phân vùng vào các thùng (tần số bằng nhau)
  - thì người ta có thể làm mịn bằng phương tiện bin, làm mịn bằng trung vị bin, làm mịn bằng ranh giới bin
- Hồi quy
  - làm trơn tru bằng cách khớp dữ liệu vào các hàm hồi quy
- Phân cụm

phát hiện và loại bỏ các ngoại lệ

- Kiểm tra kết hợp máy tính và con người

phát hiện các giá trị đáng ngờ và được con người kiểm tra (ví dụ: xử lý các giá trị ngoại lệ có thể xảy ra)

CHƯƠNG 2: BÀI TẬP VỀ Thuật TOÁN APRIORI, BỘ THÔ, CÂY QUYẾT ĐỊNH, BAYESIAN ngây thơ (LAPLACIAN)

Câu 1) Sử dụng Apriori

Sắp ra mắt.

Câu 2) Sử dụng Rough Set

B1: Tính INDIS(B). Có nghĩa là dựa vào B={attr1, attr2} danh sách xem có bao nhiêu loại giá trị khác nhau.

B2: Tính xấp xỉ dưới BX. Chi tiết từng phần tử trong INDIS(B), các giá trị trong mỗi phần tử đều phải có trong tệp X.

B3: Tính xấp xỉ trên BX. Tiết kiệm từng phần tử trong INDIS(B), chỉ cần tối thiểu một giá trị trong mỗi phần tử có mặt trong tệp X.

B4: Vùng Biên Giới BR = $\overline{BX} - \underline{BX}$

B5: Tính Người Ngoài B_OUT = $U - \overline{BX}$

B6: Tính hệ số chất lượng $\alpha B(X) = \frac{\underline{BX}}{\overline{BX}}$

- $\alpha B(X) = 1$ X là xấp xỉ rõ ràng đối với B
- $\alpha B(X) < 1$ X là xấp xỉ gần đúng đối với B

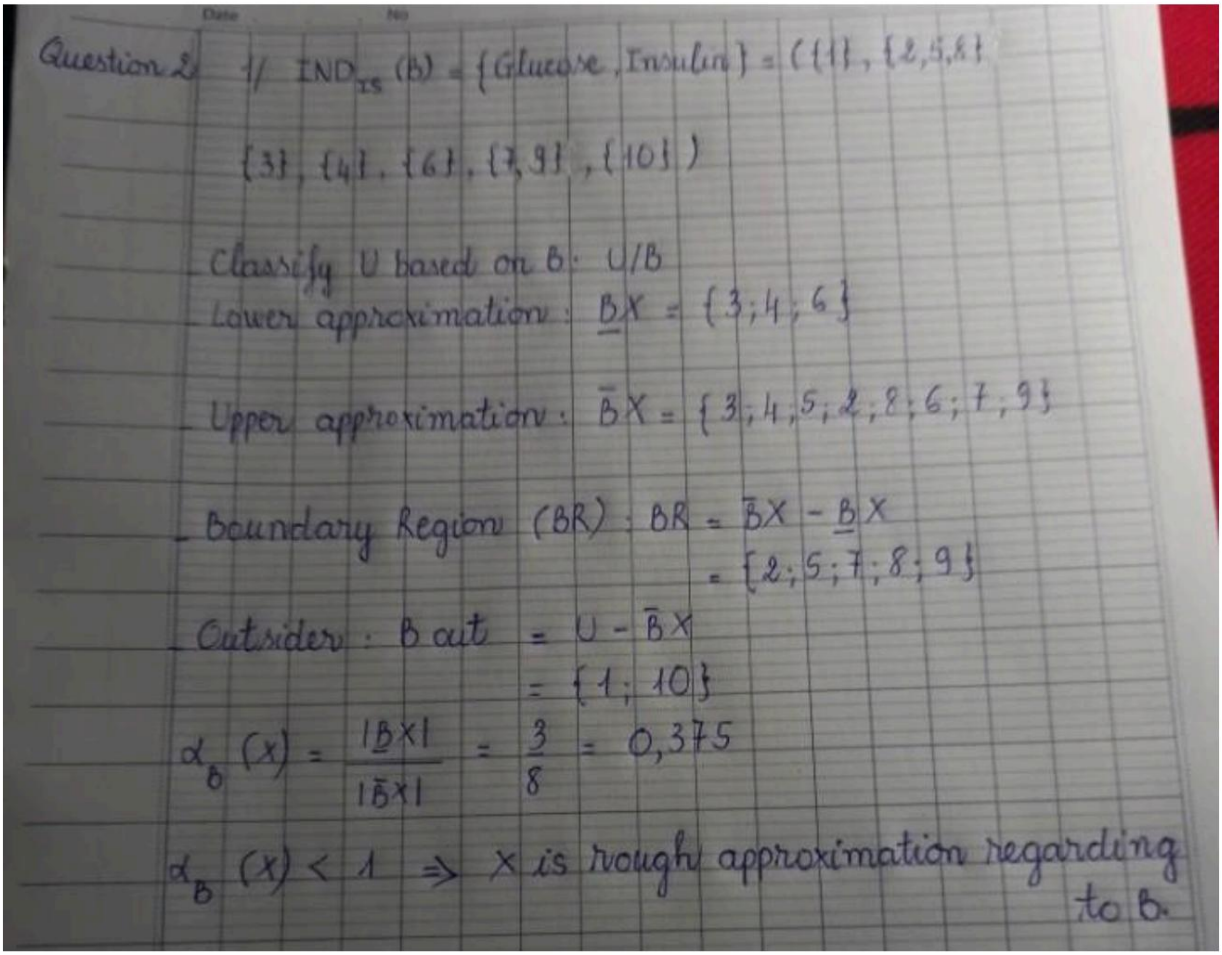
Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	Glucose (G)	Insulin (I)	BMI	Age (A)	Result (R)
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = "Positive"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*



Câu 3) Sử dụng Cây quyết định (Chỉ số Gini)

B1: Kỳ lân thuộc tính (trừ thuộc tính quyết định), xem coi là ở mỗi thuộc tính có bao

có bao nhiêu loại giá trị.

B2: Tính Gini của từng loại giá trị. Ta sẽ xem xét loại giá trị tùy thuộc vào quyết định của mình.

$$\text{Gini(SloaiGtri1)} = 1 - \left(\frac{\text{số lượng bộ thỏa mãn loại giá trị 1 và thuộc quyết định 1}}{\text{số lượng loại giá trị 1}}\right)^2 - \left(\frac{\text{số lượng loại giá trị 1 và thuộc quyết định 2}}{\text{số lượng loại giá trị 1}}\right)^2$$

Gini(SloaiGtri2) = ·

Gini(SloaiGtriN) = ·

B3: Sau khi đã tính xong Gini của từng loại giá trị. Ta sẽ tính tiếp tới Gini của thuộc tính đó.

$$\text{GiniThuốcTinh(S)} = \frac{\text{số lượng bộ của loại giá trị 1}}{\text{tổng số bộ}} * \text{Gini(SloaiGtri1)} + \frac{\text{số lượng bộ của loại giá trị 2}}{\text{tổng số bộ}} * \text{Gini(SloaiGtri2)} + \cdots$$

B4: Lặp lại các bước trên cho đến khi ko còn thuộc tính nào để xét.

B5: Chọn ra thuộc tính nào có giá trị GiniThuocTinh(S) là NHỎ NHẤT để kết luận.

"ThuốcTinh được chọn làm gốc (Chỉ số Gini là giá trị nhỏ nhất)"

Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	Glucose (G)	Insulin (I)	BMI	Age (A)	Result (R)
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = "Positive"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*
2. Determine the root of Decision Tree using Gini Index. *(2.0 scores)*

Machine Translated by Google

Date . . . No

2/ Gini Index

* Pregnancies (P)

$$Gini(S_0) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$$

$$Gini(S_1) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0,44$$

$$Gini(S_2) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \approx 0,5$$

$$Gini(S_3) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$\Rightarrow Gini_{Pregnancies}(S) = \frac{2}{10} \times 0,5 + \frac{3}{10} \times 0,44 + \frac{4}{10} \times 0,5 + \frac{1}{10} \times 0 = 0,432$$

* Glucose (G)

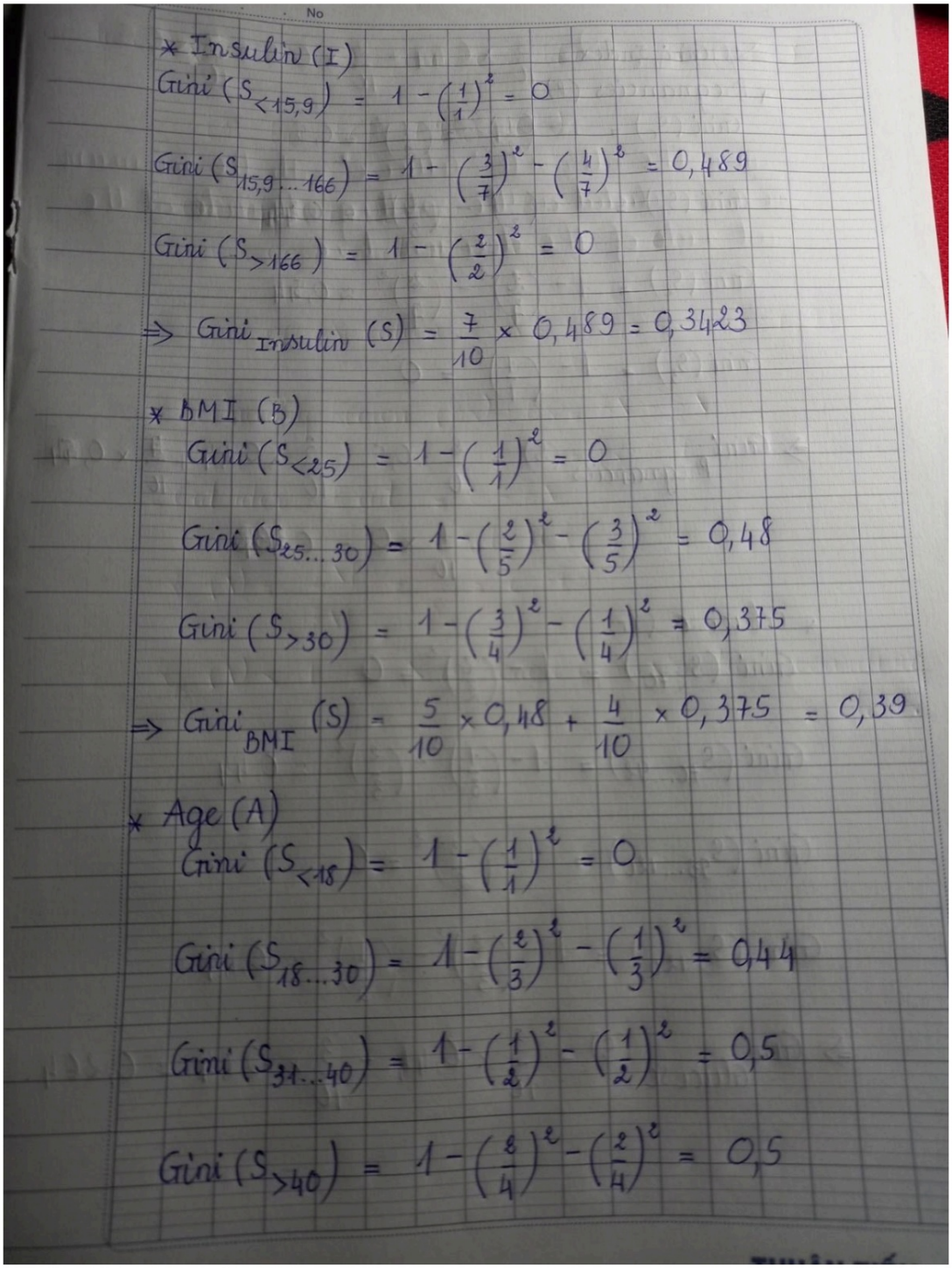
$$Gini(S_{<70}) = 1 - \left(\frac{2}{2}\right)^2 = 0$$

$$Gini(S_{70...92}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \approx 0,44$$

$$Gini(S_{93...120}) = 1 - \left(\frac{2}{2}\right)^2 = 0$$

$$Gini(S_{>120}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \approx 0,44$$

$$\Rightarrow Gini_{Glucose}(S) = \frac{3}{10} \times 0,44 + \frac{3}{10} \times 0,44 = 0,264$$



* Insulin (I)

$Gini(S_{<15,9}) = 1 - \left(\frac{1}{1}\right)^2 = 0$

$Gini(S_{15,9...166}) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0,489$

$Gini(S_{>166}) = 1 - \left(\frac{2}{2}\right)^2 = 0$

$\Rightarrow Gini_{Insulin}(S) = \frac{7}{10} \times 0,489 = 0,3423$

* BMI (B)

$Gini(S_{<25}) = 1 - \left(\frac{1}{1}\right)^2 = 0$

$Gini(S_{25...30}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0,48$

$Gini(S_{>30}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0,375$

$\Rightarrow Gini_{BMI}(S) = \frac{5}{10} \times 0,48 + \frac{4}{10} \times 0,375 = 0,39$

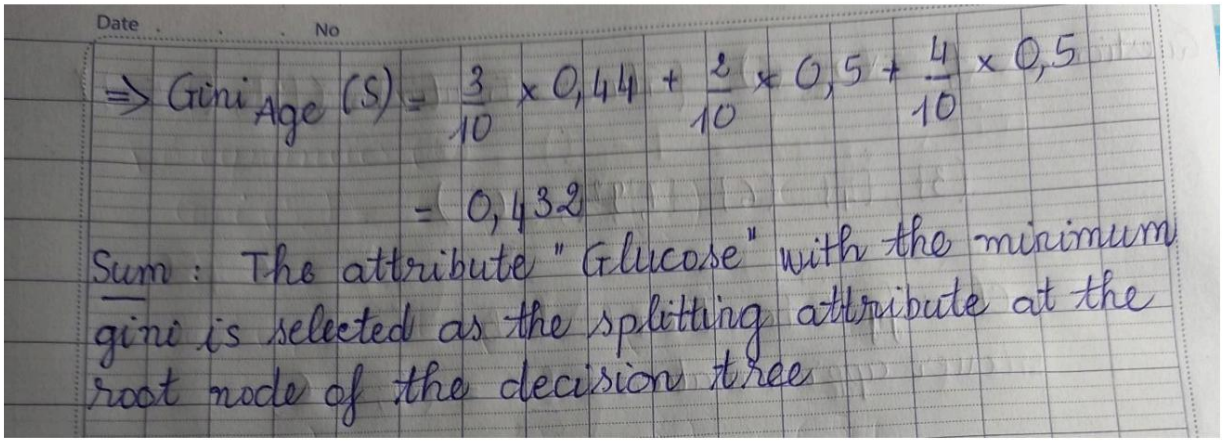
* Age (A)

$Gini(S_{<18}) = 1 - \left(\frac{1}{1}\right)^2 = 0$

$Gini(S_{18...30}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0,44$

$Gini(S_{31...40}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$

$Gini(S_{>40}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0,5$



Câu 4) Sử dụng Naive Bayesian (Laplacian)

B1: Ta chia làm hai trường hợp:

- TH1 dựa trên GIÁ TRỊ THỨ NHẤT của thuộc tính quyết định.

  vd: P(Kết quả=Dương | X) = $\frac{(\mid =)* =)}{(\ (\ )}$

- TH2 dựa trên GIÁ TRỊ CÒN LẠI của thuộc tính quyết định.

  vd: P(Kết quả=Âm tính | X) = $\frac{(\mid =)* =)}{(\ (\ )}$

B2: Tính xác thực các giá trị thuộc tính quyết định:

$$P(\text{gtriTTQD}) = \frac{\quad + 1}{+}$$

B3: Tính xác thực của X dựa trên các giá trị thuộc tính được quyết định:

$$P(X \mid \text{gtriTTQD}) = \frac{\quad +1}{+}$$

B4: Tính P(X | gtriTTQD) * P (gtriTTQD) by ta nhân gtri thu đc ở B2 với các gtri thu dc ở B3

B5: Kết luận. Ta thấy nếu P(Result=Positive | X) > P (Result=Negative | X) “There for X thuộc lớp Kết quả=Tích cực”. Và ngược lại.

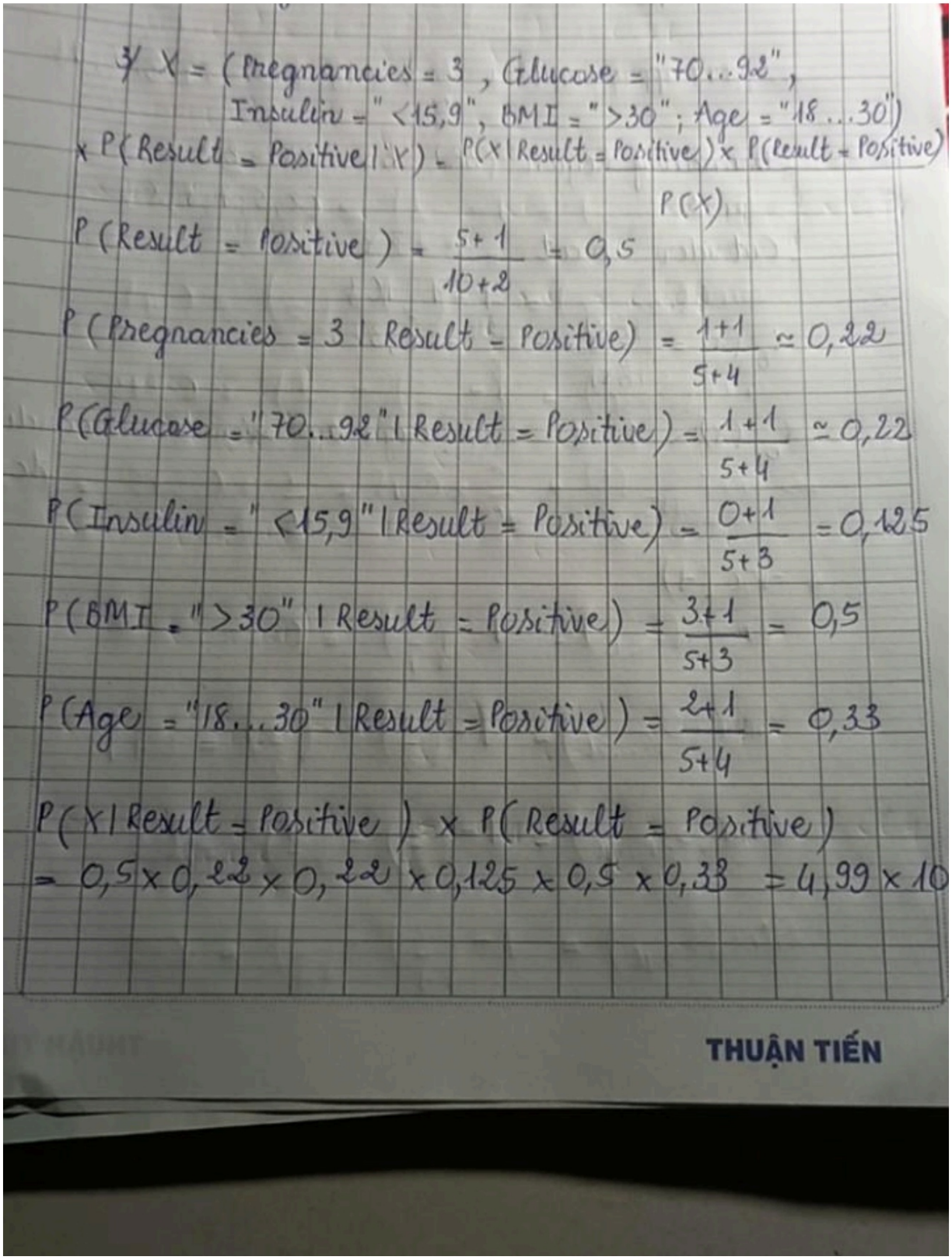
Question 2: (5.5 scores)

Suppose that a diabetes dataset as in the following table (Let *Result* be the decision attribute).

Note: Students can use abbreviations (ex. P for **Pregnancies**) to present the examination.

	Pregnancies (P)	Glucose (G)	Insulin (I)	BMI	Age (A)	Result (R)
1	0	<70	<15.9	25..30	<18	Negative
2	2	70..92	15.9 ..166	25..30	18..30	Negative
3	3	93..120	>166	>30	31..40	Positive
4	1	>120	>166	>30	>40	Positive
5	2	70..92	15.9 ..166	25..30	18..30	Positive
6	2	93..120	15.9 ..166	25..30	>40	Positive
7	0	>120	15.9 ..166	>30	18..30	Positive
8	1	70..92	15.9 ..166	<25	31..40	Negative
9	2	>120	15.9 ..166	25..30	>40	Negative
10	1	<70	15.9 ..166	>30	>40	Negative

1. Suppose B = {Glucose, Insulin}, X = {3, 4, 5, 6, 7} (***Result*** = "Positive"). Use rough set to compute: upper approximation, lower approximation, and quality coefficient *(1.5 scores)*
2. Determine the root of Decision Tree using Gini Index. *(2.0 scores)*
3. Given a sample *X = (Pregnancies = 3, Glucose = "70..92", Insulin = "<15.9", BMI = ">30"; Age = 18..30")*, what would a Naïve Bayesian classification using Laplacian correction of the ***Result*** for sample X be? *(2.0 scores)*



3/ X = (Pregnancies = 3, Glucose = "70...92",
Insulin = "<15,9", BMI = ">30"; Age = "18...30")

$$P(\text{Result} = \text{Positive} \mid X) = \frac{P(X \mid \text{Result} = \text{Positive}) \times P(\text{Result} = \text{Positive})}{P(X)}$$

$$P(\text{Result} = \text{Positive}) = \frac{5+1}{10+2} = 0,5$$

$$P(\text{Pregnancies} = 3 \mid \text{Result} = \text{Positive}) = \frac{1+1}{5+4} \simeq 0,22$$

$$P(\text{Glucose} = "70...92" \mid \text{Result} = \text{Positive}) = \frac{1+1}{5+4} \simeq 0,22$$

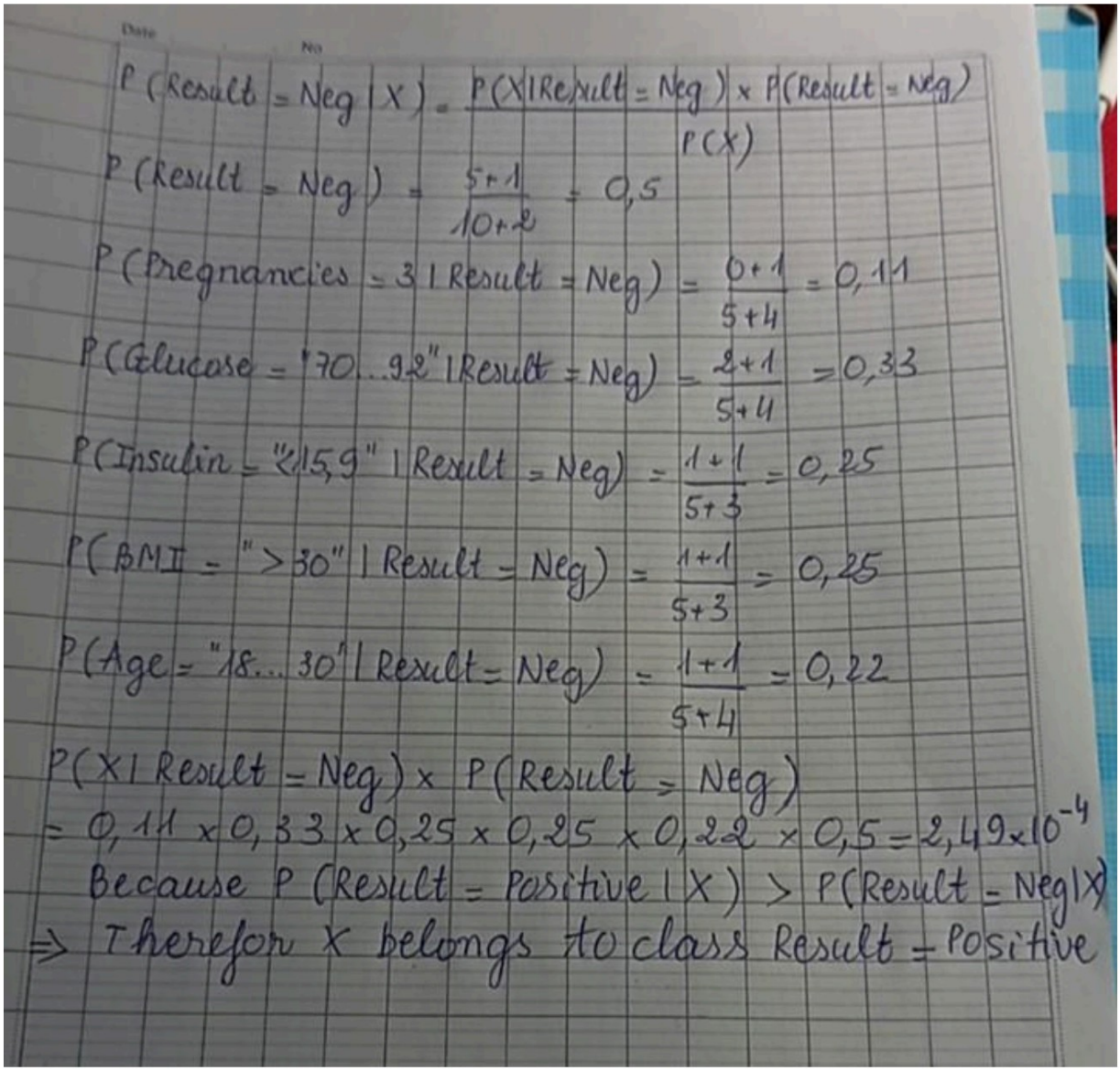
$$P(\text{Insulin} = "<15,9" \mid \text{Result} = \text{Positive}) = \frac{0+1}{5+3} = 0,125$$

$$P(\text{BMI} = ">30" \mid \text{Result} = \text{Positive}) = \frac{3+1}{5+3} = 0,5$$

$$P(\text{Age} = "18...30" \mid \text{Result} = \text{Positive}) = \frac{2+1}{5+4} = 0,33$$

$$P(X \mid \text{Result} = \text{Positive}) \times P(\text{Result} = \text{Positive})$$
$$= 0,5 \times 0,22 \times 0,22 \times 0,125 \times 0,5 \times 0,33 = 4,99 \times 10$$

THUẬN TIẾN



$$P(Result = Neg \mid X) = \frac{P(X \mid Result = Neg) \times P(Result = Neg)}{P(X)}$$

$$P(Result = Neg) = \frac{5+1}{10+2} = 0,5$$

$$P(Pregnancies = 3 \mid Result = Neg) = \frac{0+1}{5+4} = 0,11$$

$$P(Glucose = "70...92" \mid Result = Neg) = \frac{2+1}{5+4} = 0,33$$

$$P(Insulin = "\geq 15,9" \mid Result = Neg) = \frac{1+1}{5+3} = 0,25$$

$$P(BMI = ">30" \mid Result = Neg) = \frac{1+1}{5+3} = 0,25$$

$$P(Age = "18...30" \mid Result = Neg) = \frac{1+1}{5+4} = 0,22$$

$$P(X \mid Result = Neg) \times P(Result = Neg)$$

$$= 0,11 \times 0,33 \times 0,25 \times 0,25 \times 0,22 \times 0,5 = 2,49 \times 10^{-4}$$

Because $P(Result = Positive \mid X) > P(Result = Neg \mid X)$

$\Rightarrow$ Therefor X belongs to class Result = Positive

CHƯƠNG 3: BÀI TẬP VỀ THUẬT TOÁN K-MEANS

B1: Tính trung tâm Vector (Vecto tâm tâm). Ta nhìn vào cái bảng mà đề cho, ta xét từng dòng Ci, nhìn vào những ô có đánh số "1". Ta sẽ tính trung bình cộng những độ được đánh số "1".

B2: Tính khoảng cách từ các tốc độ đến từng cụm Ci. Sử dụng công thức Euclid:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

B3: Ta kẻ một bảng, cột lần đầu tiên như là các cụm Ci, dòng đầu tiên sẽ là tốc độ của các điểm.

B4: Ta sẽ so sánh khoảng cách từng bước đến các cụm Ci. Gtri nào nhỏ nhất thì cái thảo độ mà ta đang xét sẽ thuộc về cụm tương ứng. Và ta sẽ đánh số "1" tại độ đó, những cụm còn lại sẽ đánh số 0.

B5: Điểm cuối của dấu hiệu

- If like cái bảng mà ta vẽ ở B4 trùng với cái bảng mà đề cho Stop End
  end bài.
- Nếu như đề yêu cầu "Hiển thị ma trận U1" thì ta vẽ xong bảng ở B4 Stop
  Bài viết cuối cùng.
- If topic ko yêu cầu gì hết thì sau khi hoàn thành B4 mà cái bảng ta làm ko trùng vs đề bài Quay
  lại B1, làm lại các bước B1 B2 B3 B4 dựa trên cái bảng ta vừa vẽ
  Lặp lại cho đến khi bảng vẽ trùng lặp nào với bài thì thôi.

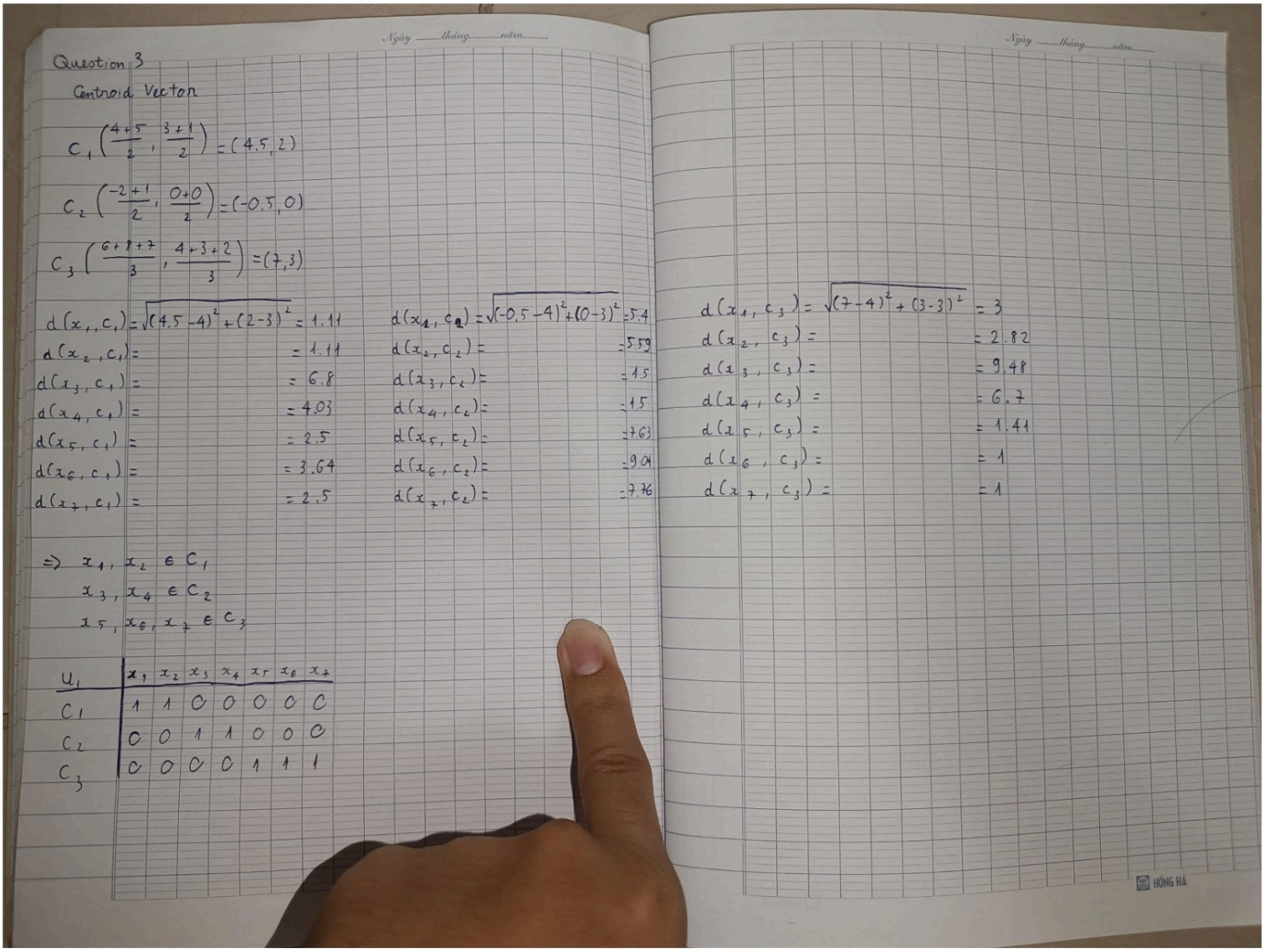
Question 3: (2.5 scores)

Suppose that 7 points as: x_1={4; 3}, x_2={5; 1}, x_3={-2; 0}, x_4={1; 0}, x_5={6; 4}, x_6={8; 3}, x_7={7; 2}, and matrix M_0 is:

M_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C_1	1	1	0	0	0	0	0
C_2	0	0	1	1	0	0	0
C3	0	0	0	0	1	1	1

Cluster the data of 7 points to 3 cluters using K_means algorithm.

Machine Translated by Google



Question 3

Centroid Vector

$C_1\left(\frac{4+5}{2}, \frac{3+1}{2}\right) = (4.5, 2)$

$C_2\left(\frac{-2+1}{2}, \frac{0+0}{2}\right) = (-0.5, 0)$

$C_3\left(\frac{6+8+7}{3}, \frac{4+3+2}{3}\right) = (7, 3)$

$d(x_1, c_1) = \sqrt{(4.5-4)^2 + (2-3)^2} = 1.11$

$d(x_2, c_1) = 1.11$

$d(x_3, c_1) = 6.8$

$d(x_4, c_1) = 4.03$

$d(x_5, c_1) = 2.5$

$d(x_6, c_1) = 3.64$

$d(x_7, c_1) = 2.5$

$d(x_1, c_2) = \sqrt{(-0.5-4)^2 + (0-3)^2} = 5.4$

$d(x_2, c_2) = 5.59$

$d(x_3, c_2) = 1.5$

$d(x_4, c_2) = 1.5$

$d(x_5, c_2) = 7.63$

$d(x_6, c_2) = 9.04$

$d(x_7, c_2) = 7.76$

$d(x_1, c_3) = \sqrt{(7-4)^2 + (3-3)^2} = 3$

$d(x_2, c_3) = 2.82$

$d(x_3, c_3) = 9.48$

$d(x_4, c_3) = 6.7$

$d(x_5, c_3) = 1.41$

$d(x_6, c_3) = 1$

$d(x_7, c_3) = 1$

$\Rightarrow x_1, x_2 \in C_1$

$x_3, x_4 \in C_2$

$x_5, x_6, x_7 \in C_3$

U_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C_1	1	1	0	0	0	0	0
C_2	0	0	1	1	0	0	0
C_3	0	0	0	0	1	1	1