🌈 Câu chuyện kỳ diệu về "Bé Na và Vương Quốc Chia Nhóm" (KMeans & KNN)

PHẦN 1: PHÉP THUẬT K-MEANS CLUSTERING

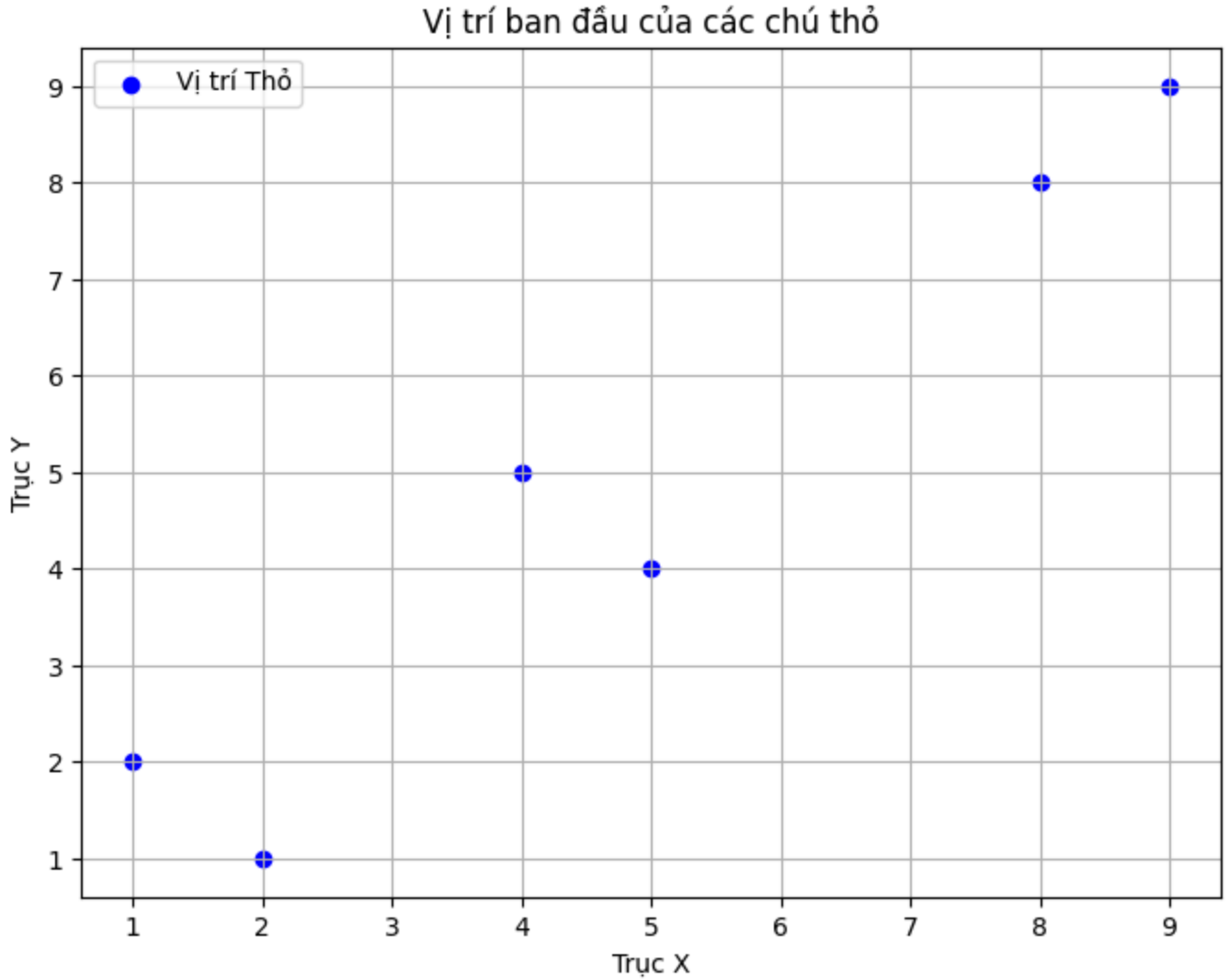
Ngày xửa ngày xưa, ở một vương quốc xa xôi, có rất nhiều chú thỏ sống rải rác khắp cánh đồng. Mỗi chú thỏ đều có một ngôi nhà nhỏ, và Bé Na – cô bé thích làm nhà khoa học – muốn giúp các chú thỏ tìm về từng nhóm bạn cùng gần nhau nhất để chơi đùa.

Đầu tiên, Bé Na chọn số lượng nhóm:

Giả sử có **K = 2** nhóm bạn (cluster).

Có tổng cộng **N = 6** chú thỏ, mỗi chú có vị trí trên bản đồ như sau:

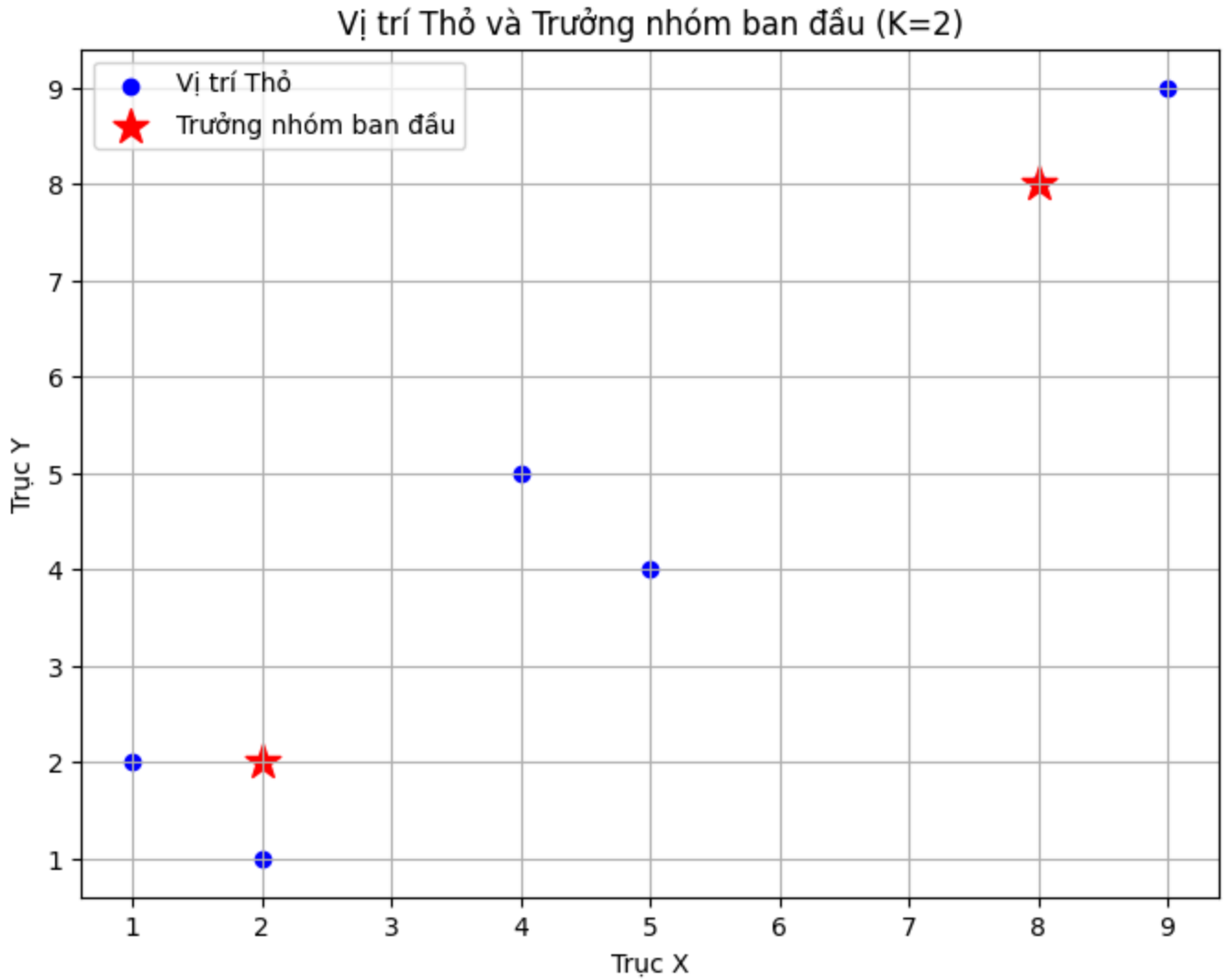
$$\((1,2),\ (2,1),\ (4,5),\ (5,4),\ (8,8),\ (9,9)\)$$


Vị trí ban đầu của các chú thỏ
Vị trí Thỏ
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

Mục tiêu của Bé Na là tìm ra cách chia các chú thỏ sao cho:
Các bạn thỏ gần nhau nhất sẽ vào cùng một nhóm, và mỗi nhóm có một "trưởng nhóm" (center) đứng ở giữa.

Các ký hiệu toán học Bé Na dùng là:

- $\(X = [x_1, x_2, \ldots, x_N]\)$$: danh sá chvị trí thỏ


Vị trí Thỏ và Trưởng nhóm ban đầu (K=2)
Vị trí Thỏ
Trưởng nhóm ban đầu
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

Để biết chú thỏ nào nên thuộc nhóm nào, Bé Na dùng công thức "gán nhãn" (one-hot):

$$y_{ik} \in \{0, 1\}, \quad \sum_{k=1}^{K} y_{ik} = 1$$

Nghĩa là mỗi chú thỏ chỉ được ở đúng một nhóm thôi!

Bé Na muốn mọi chú thỏ đều gần trưởng nhóm của mình nhất. Cô dùng một phép toán kỳ diệu gọi là **hàm mất mát**:

$$\mathcal{L}(\mathbf{Y}, \mathbf{M}) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2$$

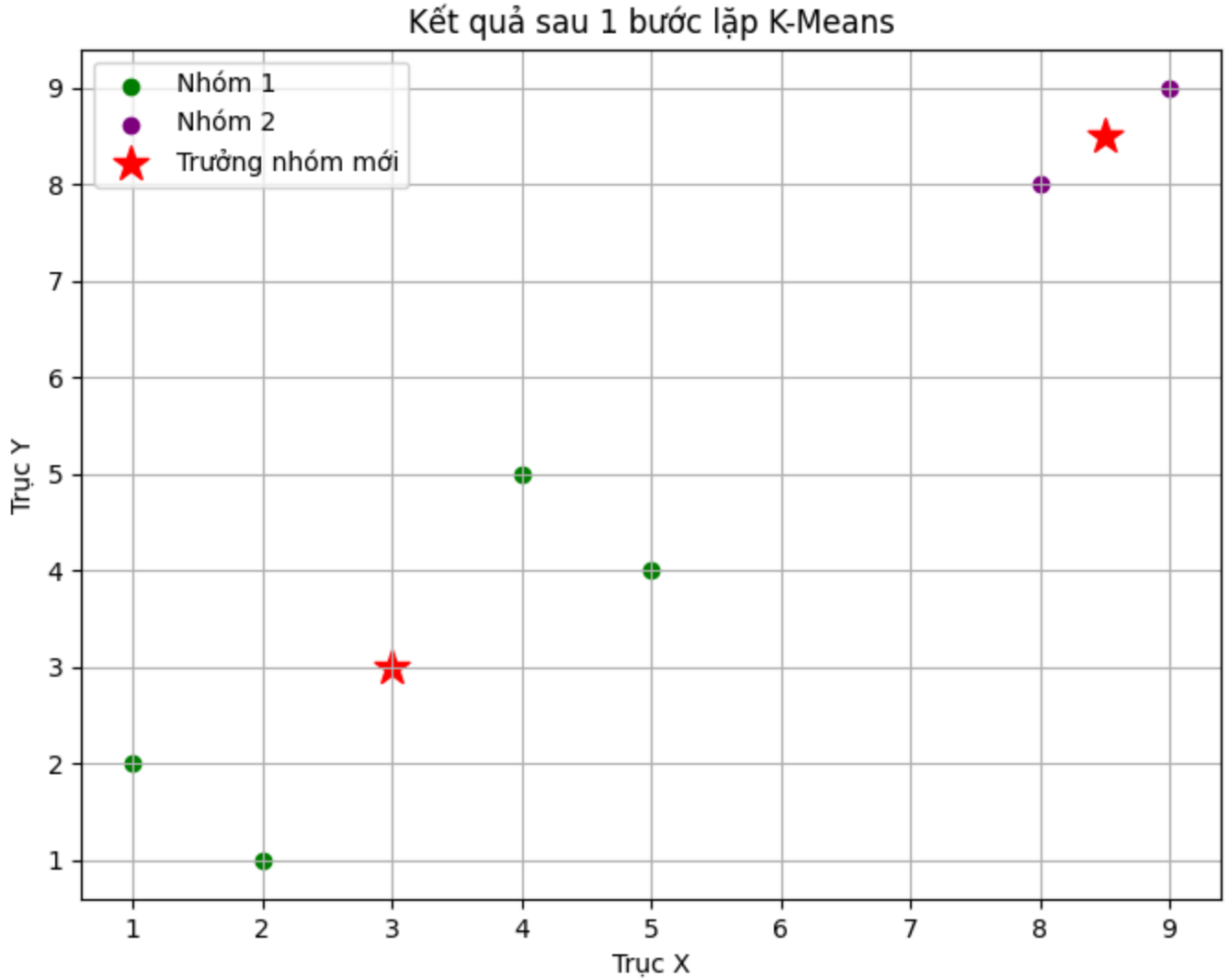
Cô phải tìm cách giảm giá trị này nhỏ nhất có thể!

Bé Na bắt đầu bằng việc chọn đại hai trưởng nhóm tạm thời, ví dụ (m_1 = (2,2)), (m_2 = (8,8)).

Bây giờ, cô xét từng chú thỏ, xem chú nào gần trưởng nhóm nào nhất, gắn nhãn cho chú đó:

$$j^* = \arg\min_j \|\mathbf{x}_i - \mathbf{m}_j\|_2^2$$

Ví dụ: chú thỏ ở ((1,2)) sẽ gần (m_1) hơn.


Kết quả sau 1 bước lặp K-Means
Nhóm 1
Nhóm 2
Trưởng nhóm mới
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

Sau khi gán xong, Bé Na cập nhật lại vị trí trưởng nhóm bằng cách lấy trung bình cộng các vị trí của các bạn thỏ trong nhóm:

$$\mathbf{m}_j = \frac{\sum_{i=1}^{N} y_{ij}\mathbf{x}_i}{\sum_{i=1}^{N} y_{ij}}$$

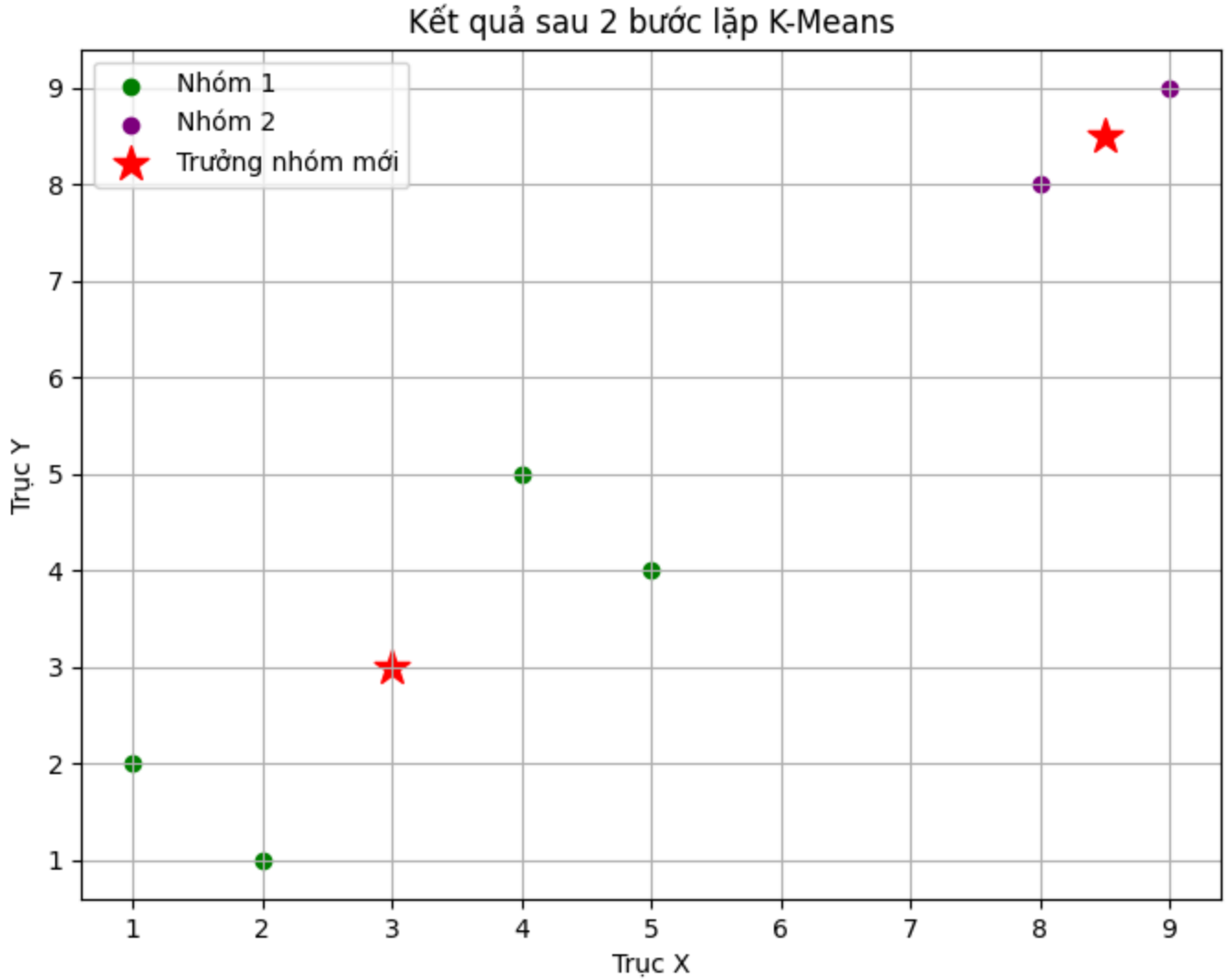
Nếu nhóm 1 có các chú ở ((1,2), (2,1), (4,5), (5,4)), trung bình cộng là:

$$\mathbf{m}_1 = \frac{(1,2) + (2,1) + (4,5) + (5,4)}{4} = (3,3)$$

Cứ lặp đi lặp lại hai bước:

1. Gán nhãn theo trưởng nhóm gần nhất
2. Cập nhật vị trí trưởng nhóm

Cho đến khi các nhóm không còn thay đổi nữa — đó là lúc mọi chú thỏ đều ở đúng nhóm bạn thân mình rồi nhé!


Kết quả sau 2 bước lặp K-Means
Nhóm 1
Nhóm 2
Trưởng nhóm mới
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

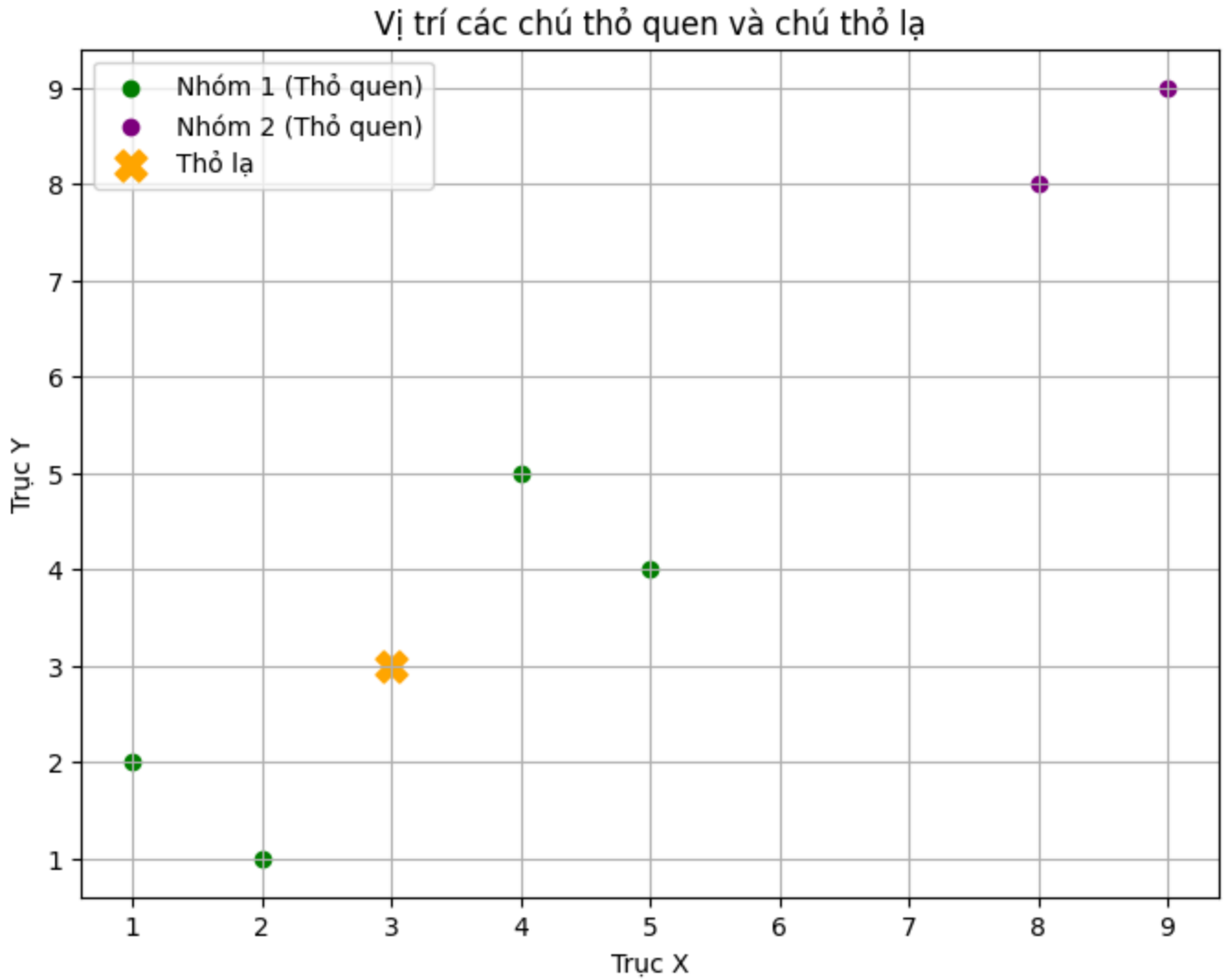
PHẦN 2: PHÉP THUẬT K-NEAREST NEIGHBORS (KNN)

Xong chuyện chia nhóm, Bé Na lại có trò mới: Khi một chú thỏ lạ xuất hiện trên cánh đồng, làm sao biết chú ấy thuộc nhóm nào?

Bé Na nghĩ ra một cách rất vui:
Mỗi khi một chú thỏ mới đến, Bé Na sẽ tìm **K** chú thỏ đã quen gần nhất, xem nhóm của họ là gì, rồi… đa số thắng!

Công thức tính khoảng cách Bé Na dùng là:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{(x_1 - x_{i1})^2 + (x_2 - x_{i2})^2}$$


Vị trí các chú thỏ quen và chú thỏ lạ
Nhóm 1 (Thỏ quen)
Nhóm 2 (Thỏ quen)
Thỏ lạ
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

Chú thỏ lạ ở vị trí: [3 3]

Chỉ số của 3 chú thỏ quen gần nhất: [0 1 2]

Vị trí của 3 chú thỏ quen gần nhất: [[1 2]

[2 1]

[4 5]]

Nhóm của 3 chú thỏ quen gần nhất: [0 0 0]

Nhóm được dự đoán cho chú thỏ lạ (dựa trên đa số): Nhóm 1

Giả sử một chú thỏ mới ở vị trí ((3,3)). Bé Na tính khoảng cách đến các bạn thỏ quen:

$$d((3,3),(1,2)) = \sqrt{(3-1)^2 + (3-2)^2} = \sqrt{4+1} = \sqrt{5}$$

$$d((3,3),(2,1)) = \sqrt{1+4} = \sqrt{5}$$

$$d((3,3),(4,5)) = \sqrt{1+4} = \sqrt{5}$$

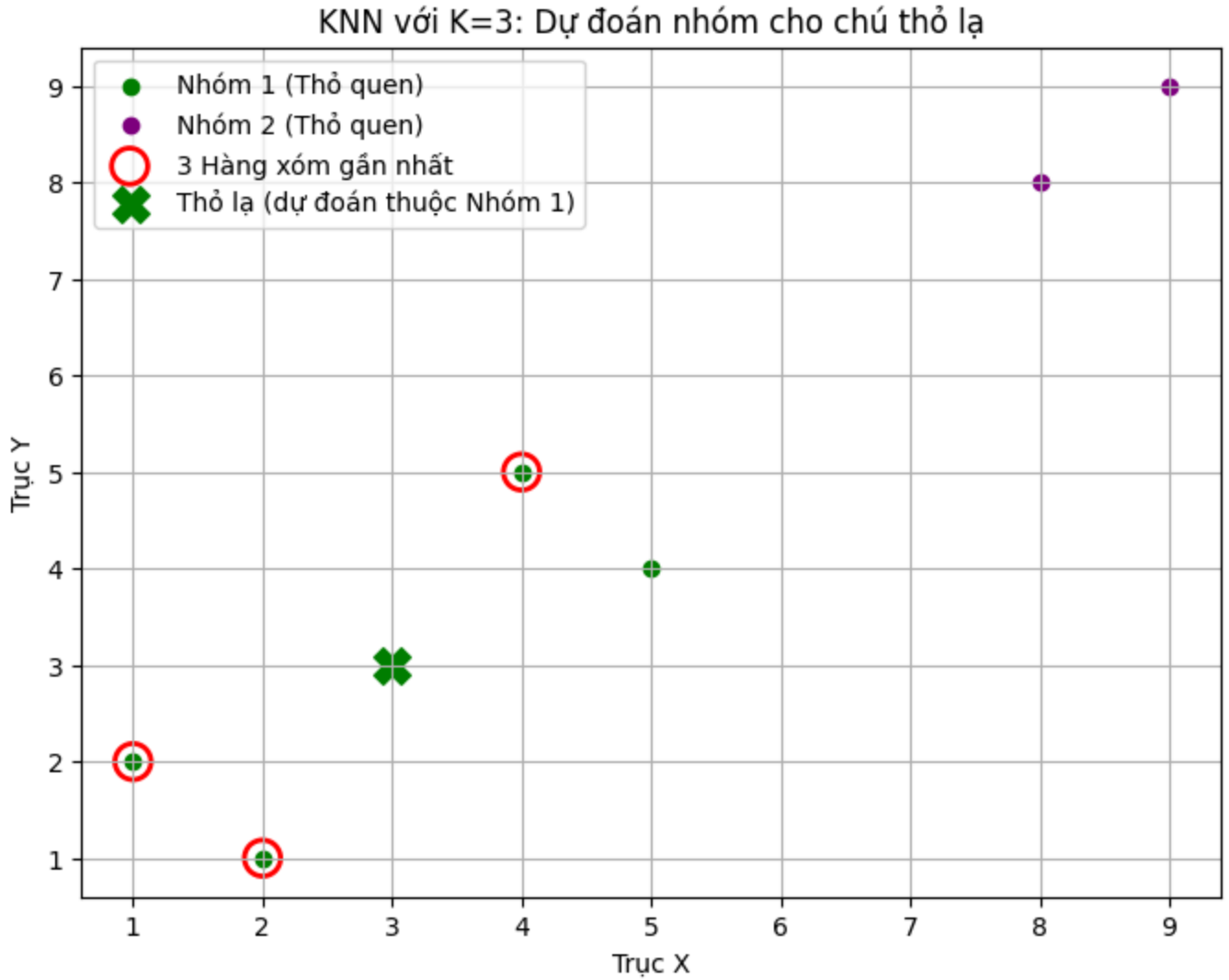
$$d((3,3),(5,4)) = \sqrt{4+1} = \sqrt{5}$$

$$d((3,3),(8,8)) = \sqrt{25+25} = \sqrt{50}$$

$$d((3,3),(9,9)) = \sqrt{36+36} = \sqrt{72}$$

Nếu chọn **K = 3**, ba bạn gần nhất là ((1,2), (2,1), (4,5)).

Giả sử ba bạn này đều thuộc nhóm 1, vậy chú thỏ mới cũng sẽ được Bé Na đưa vào nhóm 1!


KNN với K=3: Dự đoán nhóm cho chú thỏ lạ
Nhóm 1 (Thỏ quen)
Nhóm 2 (Thỏ quen)
3 Hàng xóm gần nhất
Thỏ lạ (dự đoán thuộc Nhóm 1)
Trục Y
Trục X
1
2
3
4
5
6
7
8
9

Nếu Bé Na muốn công bằng hơn, cô có thể cho điểm số các bạn gần hơn nhiều phiếu hơn, ví dụ:

$$w_i = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)}$$

Những bạn nào càng gần, phiếu càng to, bạn nào ở xa thì phiếu nhỏ thôi!

Và thế là, nhờ hai phép thuật KMeans để chia nhóm, và KNN để nhận bạn mới, Bé Na giúp vương quốc thỏ luôn vui vẻ, các bạn luôn được ở bên bạn bè thân thiết nhất của mình!

Tóm tắt thần kỳ

- KMeans: Chia nhóm các bạn dựa trên vị trí, tính toán trung bình cộng để tìm trưởng nhóm mới.
- KNN: Khi có bạn mới, nhìn những bạn gần nhất để quyết định nhóm.
- Toán học ở đây là các công thức tính khoảng cách, trung bình cộng và gán nhãn thông minh.

Và đó là câu chuyện về Bé Na, người bạn nhỏ với bộ não toán học kỳ diệu, đã giúp mọi chú thỏ tìm được mái nhà của mình!