**Chương 9. Combination of classifiers (Sự kết hợp của các bộ phân loại)**

**Outline**

[**I. CÁC CÂU HỎI CHƯƠNG 9 CÓ TRONG CÁC ĐỀ THI 2**](#_43e9l3bvebs3)

[1. Fin\_201.pdf\_HKI\_(2020-2021) 2](#_ghd6d3hz7wos)

[2. Fin\_202.pdf\_HKII\_(2020-2021) 2](#_3tl6br8wrpcf)

[3. Fin\_12\_2021.pdf\_HKI\_(2021-2022) 2](#_j9gq0wo6z7xt)

[4. Fin\_12\_2021.pdf\_HKII\_(2021-2022) 2](#_lxbo7ly6sudw)

[**II. CÂU HỎI CHI TIẾT TRONG ĐỀ 2**](#_redpzuq8njsd)

[1. Fin\_201.pdf\_HKI\_(2020-2021) 2](#_3xx3b76v0id3)

[2. Fin\_202.pdf\_HKII\_(2020-2021) 3](#_7ojfifuqfd31)

[3. Fin\_12\_2021.pdf\_HKI\_(2021-2022) 3](#_s1im3gd9yoi)

[4. Fin\_12\_2021.pdf\_HKII\_(2021-2022) 3](#_xr5cagvendmm)

[**II. GIẢI ĐỀ…To do … 4**](#_q2m3hf5vaw7i)

## **I. CÁC CÂU HỎI CHƯƠNG 9 CÓ TRONG CÁC ĐỀ THI**

##### [Fin\_201.pdf\_HKI\_(2020-2021)](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)

* + Câu 6
  + Câu 7.a)

##### [Fin\_202.pdf\_HKII\_(2020-2021)](https://docs.google.com/document/d/1DkbvgkUJOOQcx9eNk-DsMT5MwyhNMvLE/edit#)

* + Câu 6 **(Trùng với câu 6** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**
  + Câu 7
  + Câu 8

##### [Fin\_12\_2021.pdf\_HKI\_(2021-2022)](https://docs.google.com/document/d/1FBqWzRhwBudjJDBuK2_z72pM98EndxA2/edit#heading=h.t37itj2wog4b)

* + Câu 6 **(Trùng với câu 6** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**
  + Câu 7
  + Câu 8
  + Câu 9

##### [Fin\_12\_2021.pdf\_HKII\_(2021-2022)](https://docs.google.com/document/d/1BLSqsCee6LnqI5rh2J_ZeN3whI-TOJl-VinGDjto97A/edit#)

* + Câu 1. Phần 1.2
  + Câu 6 **(Trùng với câu 6** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**
  + Câu 7 **(Trùng với câu 7** [**Fin\_12\_2021.pdf\_HKI\_(2021-2022)**](https://docs.google.com/document/d/1FBqWzRhwBudjJDBuK2_z72pM98EndxA2/edit#heading=h.t37itj2wog4b)**)**
  + Câu 8

## **II. CÂU HỎI CHI TIẾT TRONG ĐỀ**

#### **1.** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)

**Câu 6. (1.5 điểm)** Cho tập mẫu gồm 14 mẫu hai chiều như sau:

X1 = (1, 1, 1), X2 = (1, 2, 1), X3 = (2, 1, 1), X4 = (2, 1.5, 1), X5 = (3, 2, 1),

X6 = (4, 1.5, 2), X7 = (4, 2, 2), X8 = (5, 1.5, 2), X9 = (4.5, 2, 2), X10 = (4, 4, 3),

X11 = (4.5, 4, 3), X12 = (4.5, 5, 3), X13 = (4, 5, 3), X14 = (5, 5, 3),

Mỗi mẫu được biểu diễn bởi đặc trưng 1, đặc trưng 2 và nhãn lớp. Ta xây dựng một tổ hợp bộ phân lớp bằng kỹ thuật bagging như sau:

Khi tạo bộ phân lớp thứ nhất, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân

lớp này : D1 = {X2, X4, X5, X8, X9, X12, X13, X14}

Khi tạo bộ phân lớp thứ hai, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân

lớp này : D2 = { X2, X4, X5, X8, X9, X10, X11}

Khi tạo bộ phân lớp thứ ba, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân

lớp này : D3 = { X1, X3, X8, X9, X12, X13, X14}

Khi tạo bộ phân lớp thứ tư, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân lớp này : D4 = { X1, X3, X8, X9, X10, X11}

Khi tạo bộ phân lớp thứ năm, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân lớp này : D5 = { X1, X3, X6, X7, X12, X13 , X14}

Khi tạo bộ phân lớp thứ sáu, bằng cách lấy mẫu, ta tạo ra tập huấn luyện cho bộ phân lớp này : D6 = { X1, X3, X6, X7, X10, X11}

Các bộ phân lớp thành phần đều là bộ phân lớp 1-lân cận gần nhất.

Xét mẫu thử (3.5, 2.8). Bộ phân lớp tổ hợp sẽ quyết định mẫu thử thuộc lớp nào ?  
**→ Chương 9**

**Giải:**

**[Tuấn]**

Khi áp dụng bộ phân lớp thành phần lên mẫu thử (3.5, 2.8), ta cần tính toán khoảng cách giữa mẫu thử và các mẫu trong tập huấn luyện của từng bộ phân lớp thành phần. Dự đoán lớp của mẫu thử sẽ được xác định bởi lớp xuất hiện nhiều nhất trong các mẫu gần nhất.

Tiếp tục từ bước trước, ta tính khoảng cách giữa mẫu thử (3.5, 2.8) và các mẫu trong tập huấn luyện của bộ phân lớp 1 (D1):

Khoảng cách giữa (3.5, 2.8) và các mẫu trong D1:

* Khoảng cách từ (3.5, 2.8) đến X2: sqrt((1-3.5)^2 + (2-2.8)^2 + (1-1)^2) = sqrt(7.65) ≈ 2.77
* Khoảng cách từ (3.5, 2.8) đến X4: sqrt((2-3.5)^2 + (1.5-2.8)^2 + (1-1)^2) = sqrt(2.65) ≈ 1.63
* Khoảng cách từ (3.5, 2.8) đến X5: sqrt((3-3.5)^2 + (2-2.8)^2 + (1-1)^2) = sqrt(0.65) ≈ 0.81
* Khoảng cách từ (3.5, 2.8) đến X8: sqrt((5-3.5)^2 + (1.5-2.8)^2 + (2-1)^2) = sqrt(3.65) ≈ 1.91
* Khoảng cách từ (3.5, 2.8) đến X9: sqrt((4.5-3.5)^2 + (2-2.8)^2 + (2-1)^2) = sqrt(1.65) ≈ 1.28
* Khoảng cách từ (3.5, 2.8) đến X12: sqrt((4.5-3.5)^2 + (5-2.8)^2 + (3-1)^2) = sqrt(8.05) ≈ 2.84
* Khoảng cách từ (3.5, 2.8) đến X13: sqrt((4-3.5)^2 + (5-2.8)^2 + (3-1)^2) = sqrt(7.25) ≈ 2.69
* Khoảng cách từ (3.5, 2.8) đến X14: sqrt((5-3.5)^2 + (5-2.8)^2 + (3-1)^2) = sqrt(8.05) ≈ 2.84

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D1:

Lớp của X2: Lớp 1

Lớp của X4: Lớp 1

Lớp của X5: Lớp 2

Lớp của X8: Lớp 2

Lớp của X9: Lớp 2

Lớp của X12: Lớp 3

Lớp của X13: Lớp 3

Lớp của X14: Lớp 3

Ta thấy rằng trong D1, có 2 mẫu gần nhất thuộc lớp 1 (X2 và X4), và 3 mẫu gần nhất thuộc lớp 3 (X12, X13, X14). Vì số lượng mẫu thuộc lớp 3 nhiều hơn, nên mẫu thử (3.5, 2.8) được dự đoán thuộc lớp 3.

Ta cần thực hiện các bước tương tự cho các bộ phân lớp thành phần khác và sau đó kết hợp kết quả của các bộ phân lớp thành phần để đưa ra dự đoán cuối cùng.

Để quyết định mẫu thử (3.5, 2.8) thuộc lớp nào bằng bộ phân lớp tổ hợp bagging, ta thực hiện các bước sau:

1. Áp dụng bộ phân lớp thành phần lên mẫu thử (3.5, 2.8):

Bộ phân lớp 1 dự đoán: X2, X4, X5, X8, X9, X12, X13, X14 -> Lớp 2

Bộ phân lớp 2 dự đoán: X2, X4, X5, X8, X9, X10, X11 -> Lớp 3

Bộ phân lớp 3 dự đoán: X1, X3, X8, X9, X12, X13, X14 -> Lớp 1

Bộ phân lớp 4 dự đoán: X1, X3, X8, X9, X10, X11 -> Lớp 1

Bộ phân lớp 5 dự đoán: X1, X3, X6, X7, X12, X13, X14 -> Lớp 3

Bộ phân lớp 6 dự đoán: X1, X3, X6, X7, X10, X11 -> Lớp 3

1. Đếm số lượng dự đoán thuộc từng lớp:

Lớp 1: 2 dự đoán

Lớp 2: 1 dự đoán

Lớp 3: 3 dự đoán

1. Chọn lớp có số lượng dự đoán nhiều nhất là kết quả dự đoán cho mẫu thử (3.5, 2.8). Trong trường hợp này, có 3 bộ phân lớp dự đoán lớp 3, nhiều nhất so với các lớp khác, vì vậy mẫu thử được quyết định thuộc lớp 3.

**[Nhân]**

Khi áp dụng bộ phân lớp thành phần lên mẫu thử (3.5, 2.8), ta cần tính toán khoảng cách giữa mẫu thử và các mẫu trong tập huấn luyện của từng bộ phân lớp thành phần. Dự đoán lớp của mẫu thử sẽ được xác định bởi lớp xuất hiện nhiều nhất trong các mẫu gần nhất.

Tiếp tục từ bước trước, ta tính khoảng cách giữa mẫu thử (3.5, 2.8) và các mẫu trong tập huấn luyện của bộ phân lớp 1 (D1):

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D1:

● Khoảng cách từ (3.5, 2.8) đến X2: sqrt((1-3.5)^2 + (2-2.8)^2) = sqrt(**6.89**) ≈ **2.62**

● Khoảng cách từ (3.5, 2.8) đến X4: sqrt((2-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.98**

● Khoảng cách từ (3.5, 2.8) đến X5: sqrt((3-3.5)^2 + (2-2.8)^2) = sqrt(**0.89**) ≈ **0.94 => Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X8: sqrt((5-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.985**

● Khoảng cách từ (3.5, 2.8) đến X9: sqrt((4.5-3.5)^2 + (2-2.8)^2) = sqrt(**1.64**) ≈ **1.281**

● Khoảng cách từ (3.5, 2.8) đến X12: sqrt((4.5-3.5)^2 + (5-2.8)^2) = sqrt(**5.84**) ≈ **2.417**

● Khoảng cách từ (3.5, 2.8) đến X13: sqrt((4-3.5)^2 + (5-2.8)^2 = sqrt(**5.09**) ≈ **2.256**

● Khoảng cách từ (3.5, 2.8) đến X14: sqrt((5-3.5)^2 + (5-2.8)^2) = sqrt(**7.09**) ≈ **2.663**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D1:

Lớp của X2: Lớp 1

Lớp của X4: Lớp 1

Lớp của X5: Lớp **1**

Lớp của X8: Lớp 2

Lớp của X9: Lớp 2

Lớp của X12: Lớp 3

Lớp của X13: Lớp 3

Lớp của X14: Lớp 3

Ta thấy rằng trong D1, có **3** mẫu gần nhất thuộc lớp 1 (X2, X4 và X5), và 3 mẫu gần nhất thuộc lớp 3 (X12, X13, X14). Vì số lượng mẫu thuộc lớp 1, 3 như nhau và mẫu thử (3.5, 2.8) có khoản cách so với X5 là ngắn nhất Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **1**.

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D2:

● Khoảng cách từ (3.5, 2.8) đến X2: sqrt((1-3.5)^2 + (2-2.8)^2) = sqrt(**6.89**) ≈ **2.62**

● Khoảng cách từ (3.5, 2.8) đến X4: sqrt((2-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.98**

● Khoảng cách từ (3.5, 2.8) đến X5: sqrt((3-3.5)^2 + (2-2.8)^2) = sqrt(**0.89**) ≈ **0.94 => Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X8: sqrt((5-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.985**

● Khoảng cách từ (3.5, 2.8) đến X9: sqrt((4.5-3.5)^2 + (2-2.8)^2) = sqrt(**1.64**) ≈ **1.281**

● Khoảng cách từ (3.5, 2.8) đến X10: sqrt((4-3.5)^2 + (4-2.8)^2) = sqrt(**1.69**) ≈ **1.300**

● Khoảng cách từ (3.5, 2.8) đến X11: sqrt((4.5-3.5)^2 + (4-2.8)^2) = sqrt(**2.44**) ≈ **1.562**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D2:

Lớp của X2: Lớp 1

Lớp của X4: Lớp 1

Lớp của X5: Lớp **1**

Lớp của X8: Lớp 2

Lớp của X9: Lớp 2

Lớp của X10: Lớp 3

Lớp của X11: Lớp 3

Ta thấy rằng trong D2, có **3** mẫu gần nhất thuộc lớp 1 (X2, X4 và X5) là nhiều nhất và mẫu thử (3.5, 2.8) có khoản cách so với X5 là ngắn nhất Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **1**.

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D3:

● Khoảng cách từ (3.5, 2.8) đến X1: sqrt((1-3.5)^2 + (1-2.8)^2) = sqrt(**9.49**) ≈ **3.08**

● Khoảng cách từ (3.5, 2.8) đến X3: sqrt((2-3.5)^2 + (1-2.8)^2) = sqrt(**5.49**) ≈ **2.34**

● Khoảng cách từ (3.5, 2.8) đến X8: sqrt((5-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.985**

● Khoảng cách từ (3.5, 2.8) đến X9: sqrt((4.5-3.5)^2 + (2-2.8)^2) = sqrt(**1.64**) ≈ **1.281**

**=> Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X12: sqrt((4.5-3.5)^2 + (5-2.8)^2) = sqrt(**5.84**) ≈ **2.417**

● Khoảng cách từ (3.5, 2.8) đến X13: sqrt((4-3.5)^2 + (5-2.8)^2) = sqrt(**5.09**) ≈ **2.256**

● Khoảng cách từ (3.5, 2.8) đến X14: sqrt((5-3.5)^2 + (5-2.8)^2) = sqrt(**7.09**) ≈ **2.663**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D3:

Lớp của X1: Lớp 1

Lớp của X3: Lớp 1

Lớp của X8: Lớp 2

Lớp của X9: Lớp 2

Lớp của X12: Lớp 3

Lớp của X13: Lớp 3

Lớp của X14: Lớp 3

Ta thấy rằng trong D3, có **2** mẫu gần nhất thuộc lớp 2 (X8 và X9), và 3 mẫu gần nhất thuộc lớp 3 (X12, X13, X14). Vì số lượng mẫu thuộc lớp 3 là nhiều nhất Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **3**.

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D4:

● Khoảng cách từ (3.5, 2.8) đến X1: sqrt((1-3.5)^2 + (1-2.8)^2) = sqrt(**9.49**) ≈ **3.08**

● Khoảng cách từ (3.5, 2.8) đến X3: sqrt((2-3.5)^2 + (1-2.8)^2) = sqrt(**5.49**) ≈ **2.34**

● Khoảng cách từ (3.5, 2.8) đến X8: sqrt((5-3.5)^2 + (1.5-2.8)^2) = sqrt(**3.94**) ≈ **1.985**

● Khoảng cách từ (3.5, 2.8) đến X9: sqrt((4.5-3.5)^2 + (2-2.8)^2) = sqrt(**1.64**) ≈ **1.281**

**=> Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X10: sqrt((4-3.5)^2 + (4-2.8)^2) = sqrt(**1.69**) ≈ **1.300**

● Khoảng cách từ (3.5, 2.8) đến X11: sqrt((4.5-3.5)^2 + (4-2.8)^2) = sqrt(**2.44**) ≈ **1.562**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D4:

Lớp của X1: Lớp 1

Lớp của X3: Lớp 1

Lớp của X8: Lớp 2

Lớp của X9: Lớp 2

Lớp của X10: Lớp 3

Ta thấy rằng trong D4, có **2** mẫu gần nhất thuộc lớp 2 (X8 và X9). Vì số lượng mẫu thuộc lớp 2 và lớp 1 như nhau và mẫu thử (3.5, 2.8) có khoản cách ngắn nhất so với X9 Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **2**.

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D5:

● Khoảng cách từ (3.5, 2.8) đến X1: sqrt((1-3.5)^2 + (1-2.8)^2) = sqrt(**9.49**) ≈ **3.08**

● Khoảng cách từ (3.5, 2.8) đến X3: sqrt((2-3.5)^2 + (1-2.8)^2) = sqrt(**5.49**) ≈ **2.34**

● Khoảng cách từ (3.5, 2.8) đến X6: sqrt((4-3.5)^2 + (1.5-2.8)^2) = sqrt(**1.94**) ≈ **1.393**

● Khoảng cách từ (3.5, 2.8) đến X7: sqrt((4-3.5)^2 + (2-2.8)^2) = sqrt(**0.89**) ≈ **0.943 => Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X12: sqrt((4.5-3.5)^2 + (5-2.8)^2) = sqrt(**5.84**) ≈ **2.417**

● Khoảng cách từ (3.5, 2.8) đến X13: sqrt((4-3.5)^2 + (5-2.8)^2) = sqrt(**5.09**) ≈ **2.256**

● Khoảng cách từ (3.5, 2.8) đến X14: sqrt((5-3.5)^2 + (5-2.8)^2) = sqrt(**7.09**) ≈ **2.663**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D5:

Lớp của X1: Lớp 1

Lớp của X3: Lớp 1

Lớp của X6: Lớp 2

Lớp của X7: Lớp 2

Lớp của X12: Lớp 3

Lớp của X13: Lớp 3

Lớp của X14: Lớp 3

Ta thấy rằng trong D5, có **3** mẫu gần nhất thuộc lớp 3 (X12, X13 và X14) và có 2 mẫu gần nhất thuộc lớp 2 (X6 và X7) Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **3**.

\* Khoảng cách giữa (3.5, 2.8) và các mẫu trong D6:

● Khoảng cách từ (3.5, 2.8) đến X1: sqrt((1-3.5)^2 + (1-2.8)^2) = sqrt(**9.49**) ≈ **3.08**

● Khoảng cách từ (3.5, 2.8) đến X3: sqrt((2-3.5)^2 + (1-2.8)^2) = sqrt(**5.49**) ≈ **2.34**

● Khoảng cách từ (3.5, 2.8) đến X6: sqrt((4-3.5)^2 + (1.5-2.8)^2) = sqrt(**1.94**) ≈ **1.393**

● Khoảng cách từ (3.5, 2.8) đến X7: sqrt((4-3.5)^2 + (2-2.8)^2) = sqrt(**0.89**) ≈ **0.943**

**=> Khoảng cách nhỏ nhất**

● Khoảng cách từ (3.5, 2.8) đến X10: sqrt((4-3.5)^2 + (4-2.8)^2) = sqrt(**1.69**) ≈ **1.300**

● Khoảng cách từ (3.5, 2.8) đến X11: sqrt((4.5-3.5)^2 + (4-2.8)^2) = sqrt(**2.44**) ≈ **1.562**

Tiếp theo, ta kiểm tra lớp của các mẫu gần nhất trong D6:

Lớp của X1: Lớp 1

Lớp của X3: Lớp 1

Lớp của X6: Lớp 2

Lớp của X7: Lớp 2

Lớp của X10: Lớp 3

Lớp của X11: Lớp 3

Ta thấy rằng trong D5, có **2** mẫu gần nhất thuộc lớp 3 (X10 và X11) và có 2 mẫu gần nhất thuộc lớp 2 (X6 và X7) và mẫu thử (3.5, 2.8) có khoản cách ngắn nhất so với X7 Nên dự đoán mẫu thử (3.5, 2.8) thuộc lớp **2**.

Để quyết định mẫu thử (3.5, 2.8) thuộc lớp nào bằng bộ phân lớp tổ hợp bagging, ta thực hiện các bước sau:

1. Áp dụng bộ phân lớp thành phần lên mẫu thử (3.5, 2.8):

Bộ phân lớp 1 dự đoán: X2, X4, X5, X8, X9, X12, X13, X14 -> Lớp **1**

Bộ phân lớp 2 dự đoán: X2, X4, X5, X8, X9, X10, X11 -> Lớp **1**

Bộ phân lớp 3 dự đoán: X1, X3, X8, X9, X12, X13, X14 -> Lớp **3**

Bộ phân lớp 4 dự đoán: X1, X3, X8, X9, X10, X11 -> Lớp **2**

Bộ phân lớp 5 dự đoán: X1, X3, X6, X7, X12, X13, X14 -> Lớp **3**

Bộ phân lớp 6 dự đoán: X1, X3, X6, X7, X10, X11 -> Lớp **2**

2. Đếm số lượng dự đoán thuộc từng lớp:

Lớp 1: 2 dự đoán

Lớp 2: 2 dự đoán

Lớp 3: 2 dự đoán

3. Chọn lớp có số lượng dự đoán nhiều nhất là kết quả dự đoán cho mẫu thử (3.5, 2.8). Trong trường hợp này, có **3 bộ phân lớp dự đoán lớp 1, lớp 2, lớp 3 là như nhau, tuy nhiên mẫu thử (3.5, 2.8) đến X5 có khoảng cách ngắn nhất là 0.94. Do đó, mẫu thử được quyết định thuộc lớp 1.**

**Câu 7. (1.5 điểm)**

1. Trong bài toán phát hiện xâm nhập mạng (intrusion detection), nếu dùng phương pháp dựa vào gom cụm (clustering-based) thay vì phương pháp dựa vào phân lớp (classification-based), thì hệ thống sẽ có được những lợi điểm gì? (*0.5 điểm*)  
   **→ Chương 6** + **Chương 9**

**Giải:**

**[Tuấn]**

Khi áp dụng phương pháp dựa vào gom cụm (clustering-based) trong bài toán phát hiện xâm nhập mạng (intrusion detection), hệ thống có thể đạt được những lợi điểm sau:

* Phát hiện các hình vi xâm nhập mới: Phương pháp gom cụm có khả năng phát hiện các hình vi xâm nhập mới mà chưa được biết trước. Khi gom cụm dựa trên dữ liệu, các điểm dữ liệu có đặc trưng tương tự sẽ được gom lại thành các nhóm. Nếu xuất hiện các điểm dữ liệu không thuộc vào các nhóm đã biết, có thể đó là các hình vi xâm nhập mới mà chưa được phân loại.
* Khả năng xử lý dữ liệu không gắn nhãn: Phương pháp gom cụm không yêu cầu dữ liệu có nhãn (không cần biết trước lớp/xâm nhập). Điều này cho phép áp dụng phương pháp gom cụm vào các tình huống mà không có thông tin về các xâm nhập cụ thể, hoặc khi dữ liệu không gắn nhãn sẵn có.
* Phát hiện các nhóm xâm nhập tương tự: Bằng cách gom cụm các điểm dữ liệu, phương pháp này có thể tìm ra các nhóm có hình vi xâm nhập tương tự. Điều này giúp hiểu rõ hơn về cách thức tấn công, mô hình hóa các hành vi xâm nhập và xác định các đặc trưng chung của các nhóm xâm nhập.
* Giảm số lượng dữ liệu đào tạo: Khi áp dụng phương pháp gom cụm, ta có thể sử dụng một tập dữ liệu nhỏ hơn so với phương pháp phân lớp truyền thống. Thay vì phải có dữ liệu được gắn nhãn cho từng lớp xâm nhập cụ thể, ta chỉ cần một tập dữ liệu không gắn nhãn và áp dụng phương pháp gom cụm để xác định các nhóm xâm nhập.

Tuy nhiên, cũng cần lưu ý rằng phương pháp gom cụm cũng có nhược điểm, bao gồm khả năng gom nhầm các điểm dữ liệu không thuộc cùng một lớp xâm nhập và khó khăn trong việc xác định ngưỡng/ranh giới cho từng nhóm. Điều này có thể dẫn đến việc có thể bỏ sót các hình vi xâm nhập hoặc phát hiện sai sót.

#### **2.** [**Fin\_202.pdf\_HKII\_(2020-2021)**](https://docs.google.com/document/d/1DkbvgkUJOOQcx9eNk-DsMT5MwyhNMvLE/edit#)

**Câu 6.**  **(Trùng với câu 6** [Fin\_201.pdf\_HKI\_(2020-2021)](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**

**Câu 7 (0.5 điểm)** Nêu hai yếu tố ngẫu nhiên được đưa vào giải thuật tổ hợp phân lớp RandomForest.

**→ Chương 9**

**Giải:**

Trong giải thuật tổ hợp phân lớp RandomForest, hai yếu tố ngẫu nhiên được đưa vào để tạo sự đa dạng và tránh overfitting trong quá trình xây dựng các cây quyết định là:

1. Bootstrap Sampling: Yếu tố này đảm bảo mỗi cây quyết định trong RandomForest được xây dựng trên một tập con dữ liệu huấn luyện ngẫu nhiên. Quá trình này sử dụng phương pháp tái chọn mẫu theo bootstrap, trong đó mỗi cây được xây dựng bằng cách lấy mẫu từ tập dữ liệu gốc với việc thay thế. Điều này dẫn đến việc một số mẫu có thể xuất hiện nhiều lần trong tập con của mỗi cây, trong khi một số mẫu khác có thể không được chọn.
2. Random Feature Subsetting: Yếu tố này đảm bảo rằng mỗi cây quyết định chỉ sử dụng một số lượng ngẫu nhiên các đặc trưng để xây dựng quyết định tại mỗi nút. Thay vì sử dụng toàn bộ đặc trưng có sẵn, mỗi cây chỉ sử dụng một phần con của chúng. Số lượng đặc trưng được chọn ngẫu nhiên có thể được xác định trước hoặc được lựa chọn tự động trong quá trình xây dựng cây.

Nhờ hai yếu tố ngẫu nhiên này, RandomForest có khả năng giảm overfitting và tạo ra một tập hợp đa dạng các cây quyết định, đóng góp vào tính ổn định và khả năng tổng hợp thông tin của mô hình.

**Câu 8. (0.5 điểm)** Nêu hai cách tiếp cận chính để giải quyết vấn đề mất cân bằng dữ liệu (imbalanced data) trong bài toán phân lớp.

**→ Chương 9.**

**Giải:**

Trong bài toán phân lớp với dữ liệu mất cân bằng (imbalanced data), có hai cách tiếp cận chính để giải quyết vấn đề này là:

1. Over-sampling: Cách tiếp cận này tập trung vào tăng cường số lượng mẫu của lớp thiểu số để đạt được sự cân bằng về số lượng mẫu giữa các lớp. Các phương pháp over-sampling tạo thêm các mẫu nhân bản từ các mẫu hiện có của lớp thiểu số. Có các kỹ thuật over-sampling phổ biến như Random Over-sampling, SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling),...
2. Under-sampling: Cách tiếp cận này tập trung vào giảm số lượng mẫu của lớp đa số để đạt được sự cân bằng. Các phương pháp under-sampling loại bỏ hoặc giảm số lượng mẫu từ lớp đa số để phù hợp với số lượng mẫu của lớp thiểu số. Các kỹ thuật under-sampling phổ biến bao gồm Random Under-sampling, Tomek Links, Cluster Centroids, NearMiss, One-Sided Selection,...

Ngoài hai cách tiếp cận trên, còn có những phương pháp kết hợp cả over-sampling và under-sampling, được gọi là phương pháp kết hợp (hybrid methods), như SMOTE-ENN (SMOTE followed by Edited Nearest Neighbors), SMOTE-Tomek, và SMOTE-Borderline-SMOTE SVM.

Việc chọn cách tiếp cận phù hợp phụ thuộc vào bộ dữ liệu cụ thể và mục tiêu của bài toán. Tuy nhiên, cần lưu ý rằng việc áp dụng over-sampling có thể dẫn đến overfitting và tiêu cực hóa đặc trưng của lớp thiểu số, trong khi under-sampling có thể gây mất mát thông tin quan trọng. Điều này đòi hỏi sự cân nhắc và thử nghiệm để tìm ra phương pháp tốt nhất cho bài toán cụ thể.

#### **3.** [**Fin\_12\_2021.pdf\_HKI\_(2021-2022)**](https://docs.google.com/document/d/1FBqWzRhwBudjJDBuK2_z72pM98EndxA2/edit#heading=h.t37itj2wog4b)

**Câu 6**  **(Trùng với câu 6** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**

**Câu 7 ((1 điểm)**

1. Nêu những điểm khác biệt giữa hai phương pháp tổ hợp bộ phân lớp: bagging và boosting (*0.5 điểm*).  
     
   **[Trâm]**

**Bagging (Bootstrap Aggregating):**

* Là một phương pháp tổ hợp bộ phân lớp trong đó các mô hình con được huấn luyện độc lập và sau đó kết hợp dự đoán của chúng thông qua việc lấy trung bình (trong trường hợp của bài toán phân lớp) hoặc lấy đa số phiếu bầu (trong trường hợp của bài toán phân loại).
* Các mô hình con trong bagging có thể sử dụng cùng một thuật toán hoặc khác nhau, và dữ liệu huấn luyện của mỗi mô hình con được tạo ra thông qua một quá trình tái chọn dữ liệu ngẫu nhiên từ tập dữ liệu huấn luyện gốc bằng phương pháp tái chọn bootstrap.
* Bagging giúp giảm phương sai (variance) của mô hình, làm giảm khả năng overfitting và cải thiện tính tổng quát hóa.

**Boosting:**

* Là một phương pháp tổ hợp bộ phân lớp, tuy nhiên, khác với bagging, các mô hình con trong boosting được huấn luyện tuần tự và tập trung vào việc cải thiện các điểm dữ liệu mà các mô hình trước đó dự đoán sai.
* Trong quá trình huấn luyện, các mô hình con sau này được tạo ra bằng cách tạo ra một hàm dự đoán mới là tổng của các hàm dự đoán của các mô hình trước đó. Các trọng số được gán cho từng mô hình con được điều chỉnh sao cho các điểm dữ liệu bị dự đoán sai trước đó nhận được sự quan tâm cao hơn.
* Boosting tập trung vào việc cải thiện hiệu suất của mô hình trên các điểm dữ liệu khó khăn, có trọng số cao hơn, và thường dẫn đến các mô hình có độ chính xác cao hơn trên tập dữ liệu kiểm tra.

1. Nêu hai yếu tố ngẫu nhiên được đưa vào giải thuật tổ hợp phân lớp RandomForest (*0.5 điểm*).  
    **→ Chương 9**

**Câu 8. (0.5 điểm)** Nêu hai cách tiếp cận chính để giải quyết vấn đề mất cân bằng dữ liệu (imbalanced data) trong bài toán phân lớp.  
 **→ Chương 9**

**Câu 9.** (**0.5 điểm**) Trong bài toán phát hiện xâm nhập mạng (intrusion detection), nếu dùng phương pháp dựa vào gom cụm (clustering-based) thay vì phương pháp dựa vào phân lớp (classification-based), thì hệ thống sẽ có được những lợi điểm gì?  
 **→ Chương 6 + Chương 9**

#### **4.** [**Fin\_12\_2021.pdf\_HKII\_(2021-2022)**](https://docs.google.com/document/d/1BLSqsCee6LnqI5rh2J_ZeN3whI-TOJl-VinGDjto97A/edit#)

**Câu 1**

**1.2** Phát biểu nào sau đây là không đúng  
A. Bài toán *quy hoạch toàn phương* (Quadratic Program

ming) liên quan đến máy vector hỗ trợ thuộc loại tối ưu hóa có ràng buộc (constrained optimization). *→ Chương 8*  
B. Trong giải thuật ADABOOST, việc xây dựng các bộ phận lớp thành phần là độc lập với nhau.*→ Chương 9*  
C. Bagging và Rừng Ngẫu nhiên sử dụng cùng chiến lược phiếu bầu: đa số phiếu (majority voting).*→ Chương 9*  
D. Quá trình huấn luyện mạng nơ ron truyền thẳng thường nhanh hơn quá trình huấn luyện mạng nơ ron RBF.*→ Chương 7*  
E. Gibb Sampling là một kỹ thuật toán học được dùng để huấn luyện mạng Deep Belief network mà dựa vào máy Boltzmann giới hạn. *→ Chương 12*  
F. Học không có giám sát không cần đến thông tin nhân lớp của các mẫu trong tập dữ liệu huấn luyện. *→ Chương 6*  (0.5 điểm)

**→ Chương 6 + Chương 7 + Chương 8 + Chương 9 + Chương 12**

**Giải:**

Phát biểu B là không đúng. Trong giải thuật ADABOOST, việc xây dựng các bộ phân lớp thành phần không độc lập với nhau. Các bộ phân lớp thành phần trong ADABOOST được xây dựng tuần tự, với mỗi bộ phân lớp sau đó được điều chỉnh để tập trung vào việc phân loại các mẫu bị phân loại sai bởi các bộ phân lớp trước đó. Quá trình này dựa trên việc điều chỉnh trọng số của các mẫu trong quá trình huấn luyện để tạo ra mô hình tổ hợp mạnh hơn.

Phát biểu C là không đúng. Bagging và Rừng Ngẫu nhiên (Random Forest) sử dụng chiến lược phiếu bầu khác nhau, không phải là majority voting.

**Câu 6 (Trùng với câu 6** [**Fin\_201.pdf\_HKI\_(2020-2021)**](https://docs.google.com/document/d/1O30YVMcYqIh3FvXyK30TFXqQB4cgPsMO/edit#heading=h.k859l5ku3md)**)**

**Câu 7.** **(Trùng với câu 7** [**Fin\_12\_2021.pdf\_HKI\_(2021-2022)**](https://docs.google.com/document/d/1FBqWzRhwBudjJDBuK2_z72pM98EndxA2/edit#heading=h.t37itj2wog4b)**)**

Câu 8. **(0.75 pt)**

a. Đường cong ROC là gì? (0.5 điểm)

b. Làm cách nào sử dụng đường cong ROC để đánh giá chất lượng của một bộ phân lớp? (0.25 điểm)  
 **→ Chương 9**

Giải:

a. Đường cong ROC (Receiver Operating Characteristic) là một công cụ đánh giá hiệu suất của một bộ phân lớp trong các bài toán phân loại. Nó biểu thị mối quan hệ giữa tỷ lệ True Positive (TPR) và tỷ lệ False Positive (FPR) trên toàn bộ các ngưỡng quyết định khác nhau.

Đường cong ROC được biểu diễn trên một biểu đồ hai trục, với trục hoành biểu thị tỷ lệ FPR và trục tung biểu thị tỷ lệ TPR. Đường cong ROC mô tả sự thay đổi của TPR và FPR khi ngưỡng quyết định thay đổi từ một giá trị đến giá trị khác.

b. Để đánh giá chất lượng của một bộ phân lớp bằng đường cong ROC, ta có thể sử dụng các đánh giá dựa trên đường cong ROC như sau:

1. Diện tích dưới đường cong ROC (AUC-ROC): Đây là một phép đo quan trọng của đường cong ROC, biểu thị khả năng phân biệt của bộ phân lớp. Giá trị AUC-ROC nằm trong khoảng từ 0 đến 1, với giá trị càng gần 1 thể hiện khả năng phân loại tốt hơn. Một bộ phân lớp với AUC-ROC lớn hơn 0.5 được coi là có khả năng phân loại tốt hơn ngẫu nhiên.
2. Điểm cắt (Threshold): Đường cong ROC cung cấp thông tin về hiệu suất của bộ phân lớp ở các ngưỡng quyết định khác nhau. Điểm cắt có thể được chọn để đáp ứng yêu cầu cụ thể của bài toán, chẳng hạn như tối ưu độ nhạy (sensitivity) hoặc độ đặc hiệu (specificity).
3. Đánh giá so sánh: Đường cong ROC cho phép so sánh hiệu suất giữa các bộ phân lớp khác nhau. Bộ phân lớp có đường cong ROC nằm trên phía trên và bên trái của bộ phân lớp khác được coi là có hiệu suất tốt hơn.

Sử dụng các phép đo và thông tin từ đường cong ROC, ta có thể đánh giá chất lượng của bộ phân lớp và chọn bộ phân lớp phù hợp cho bài toán phân loại.

## **II. GIẢI ĐỀ**…To do …