**Chương 6. Unsupervised Learning – Clustering**

**Outline**

[**I. Các câu hỏi chương 6 có trong các đề thi 2**](#_43e9l3bvebs3)

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

[2. Fin\_202.pdf\_HKII\_(2020-2021) → Không có chương 6 trong đề thi này 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)

[Câu 2. (0.5 điểm) 2](#_t5larst0hyd6)

[Câu 7. (1.5 điểm) 2](#_xy0djk5nnckx)

[2. Fin\_202.pdf\_HKII\_(2020-2021) → Không có chương 6 trong đề thi này 3](#_7ojfifuqfd31)

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

[Câu 9. (0.5 điểm) 3](#_2z045wa2v1om)

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

[Câu 1. 4](#_xa0j82mzkbzf)

[**II. Giải đề …To do … 4**](#_q2m3hf5vaw7i)

## **I. Các câu hỏi chương 6 có trong các đề thi**

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

* 1. Câu 2
  2. Câu 7

##### [Fin\_202.pdf\_HKII\_(2020-2021)](https://docs.google.com/document/d/1DkbvgkUJOOQcx9eNk-DsMT5MwyhNMvLE/edit#) **→ Không có chương 6 trong đề thi này**

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

* 1. Câu 9

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

* 1. Câu 1. (Phần 1.2)

## **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 2. (0.5 điểm)**

Các tiêu chí đánh giá ngoại như Rand, Jackard, v.v. để đánh giá chất lượng gom cụm được áp dụng trong trường hợp nào của gom cụm?

**[HXLong]** [Xác nhận đúng - Thành]

· Rand Index (RI): RI đo lường tỷ lệ giữa số lượng cặp điểm dữ liệu được cùng phân loại trong cả tập nhãn thật và tập nhãn dự đoán trên tổng số lượng cặp điểm dữ liệu. RI có giá trị từ 0 (hai phân loại hoàn toàn không giống nhau) đến 1 (hai phân loại hoàn toàn giống nhau).

· Jaccard Index: Jaccard Index đo lường tỷ lệ giữa số lượng cặp điểm dữ liệu được cùng phân loại trong cả tập nhãn thật và tập nhãn dự đoán trên số lượng cặp điểm dữ liệu ít nhất được phân loại cùng nhóm (cùng nhóm bao gồm cùng tập nhãn thật => cùng lớp, hoặc cùng tập nhãn dự đoán => cùng cụm) một lần. Jaccard Index cũng có giá trị từ 0 đến 1.

**[Thành]**

Bổ sung công thức trong slide:

a = số cặp cùng thuộc lớp G và cùng gôm vào cụm A

b = số cặp cùng thuộc lớp G và không cùng gôm vào cụm A

c = số cặp không cùng thuộc lớp G và cùng gôm vào cụm A

d = số cặp không cùng thuộc lớp G và không cùng gôm vào cụm A

* Để tính tỷ lệ Rand Index: ((a + d) / (a + b + c + d) thuộc [0, 1]
* Để tính tỷ lệ Jaccard Index: (a) / (a + b + c) thuộc [0, 1]

(Tham khảo slide 47-48, chương 6)

##### **Câu 7. (1.5 điểm)**

**a.** 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*)

**[HXLong]** [Xác nhận đúng - Thành]

- Clustering có thể giúp phát hiện các cuộc tấn công mới mà chưa từng được ghi nhận trước đó.

- Clustering là một kỹ thuật học không giám sát, tức là nó không yêu cầu dữ liệu gán nhãn để huấn luyện.

- Clustering có thể xử lý dữ liệu lớn một cách hiệu quả, đặc biệt với những thuật toán như K-means hay DBSCAN.

- Kết quả của gom cụm có thể giúp nhận biết các mẫu hành vi bất thường, giúp cho việc phân tích hậu vụ và tìm hiểu hơn về cách thức hoạt động của các mối đe dọa.

**→ Chương 6** + **Chương 9**  
**b.** Hãy giải thích hai thuật ngữ trong giải thuật gom cụm BIRCH: đặc trưng cụm (cluster feature) và cây đặc trưng cụm (CF tree). (*0.75 điểm*)

[HXLong] [Xác nhận đúng - Thành]

· Đặc trưng Cụm (Cluster Feature - CF): CF giúp cung cấp thông tin tóm tắt về cụm mà không cần giữ toàn bộ dữ liệu, điều này giúp tối ưu hóa bộ nhớ và tăng tốc độ xử lý.

· Cây Đặc Trưng Cụm (CF Tree): Đây là một cấu trúc dữ liệu dạng cây được sử dụng trong thuật toán BIRCH để lưu trữ và quản lý các CF. Cây CF giúp cung cấp một cấu trúc dữ liệu hiệu quả để tìm kiếm và cập nhật các cụm trong quá trình gom cụm.

**→ Chương 6  
c.** (Đúng/Sai) BIRCH là giải thuật gom cụm gia tăng. (*0.25 điểm*)

**[HXLong]** [Xác nhận đúng - Thành]

Đúng. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) là một thuật toán gom cụm gia tăng. Nó có thể xử lý dữ liệu một cách tuần tự và cập nhật mô hình gom cụm sau mỗi điểm dữ liệu mới.

**→ Chương 6   
*(Trùng với câu 9.*** [**Fin\_12\_2021.pdf\_HKI\_(2021-2022)**](https://docs.google.com/document/d/1FBqWzRhwBudjJDBuK2_z72pM98EndxA2/edit#heading=h.t37itj2wog4b)**)**

#### **2.** [**Fin\_202.pdf\_HKII\_(2020-2021)**](https://docs.google.com/document/d/1DkbvgkUJOOQcx9eNk-DsMT5MwyhNMvLE/edit#) **→ Không có chương 6 trong đề thi này**

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

##### **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?

[HXLong]

Xem câu 7 FIN201.

**→ 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.**

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

[HXLong] [Xác nhận đúng - Thành]

F. Đúng. Học không có giám sát (Unsupervised Learning) không cần đến thông tin nhãn lớp (class labels) của các mẫu trong tập dữ liệu huấn luyện. Phương pháp này tìm cách hiểu cấu trúc của dữ liệu mà không cần nhãn dữ liệu.

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

## **II. Giải đề** …To do …

……..