**Lê Hữu Hạnh 15110041**

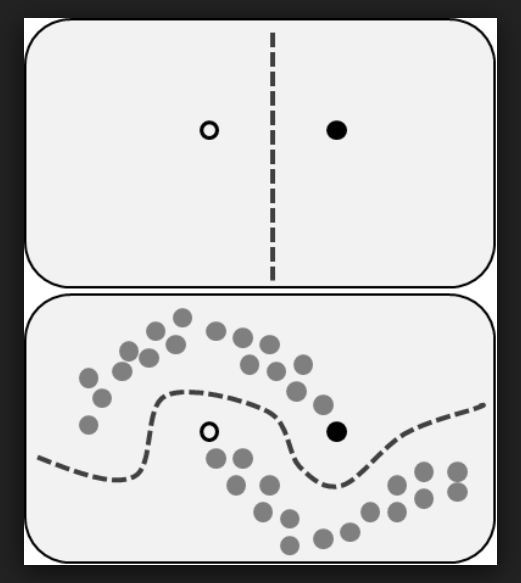
**Vũ Ngọc Phát 15110096**

**ĐỒ ÁN 3**

**TÌM HIỂU SEMI-SUPERVISED LEARNING VÀ ỨNG DỤNG**

1.Semi-supervised learning (học bán giám sát) là gì :

\_là một lớp các nhiệm vụ và kỹ thuật [học máy](https://en.wikipedia.org/wiki/Machine_learning" \o "Học máy) cũng sử dụng [dữ liệu](https://en.wikipedia.org/wiki/Data" \o "Dữ liệu) không được dán nhãn để đào tạo - thường là một lượng nhỏ [dữ liệu được dán nhãn](https://en.wikipedia.org/wiki/Labeled_data" \o "Dữ liệu được gắn nhãn) với một lượng lớn dữ liệu không được dán nhãn



\_Nằm giữa [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) (không có bất kỳ dữ liệu đào tạo được dán nhãn nào) [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) (với dữ liệu huấn luyện được dán nhãn hoàn toàn)

\_Nhiều nhà nghiên cứu học máy đã phát hiện ra rằng dữ liệu không dán nhãn, khi được sử dụng kết hợp với một lượng nhỏ dữ liệu có nhãn, có thể cải thiện đáng kể tính chính xác của việc học tập không giám sát (không có dữ liệu), nhưng không có thời gian và chi phí cần thiết cho người giám sát học tập (nơi tất cả dữ liệu được dán nhãn)

\_Việc mua lại dữ liệu có nhãn cho một vấn đề học tập thường đòi hỏi một nhân viên có tay nghề cao (ví dụ như phiên âm một đoạn âm thanh) hoặc một thí nghiệm vật lý (ví dụ: xác định cấu trúc 3D của một protein hoặc xác định xem có dầu ở một vị trí cụ thể)

\_Chi phí liên quan đến quá trình ghi nhãn do đó có thể khiến một bộ huấn luyện được dán nhãn hoàn toàn không khả thi, trong khi việc mua lại dữ liệu không có nhãn là tương đối rẻ. Trong những tình huống như vậy, việc học bán giám sát có thể có giá trị thực tiễn tuyệt vời.

2.Các giả định được sử dụng

\_Để sử dụng bất kỳ dữ liệu không được dán nhãn nào, chúng tôi phải giả định một số cấu trúc cho việc phân phối dữ liệu cơ bản. Thuật toán học bán được giám sát sử dụng ít nhất một trong các giả định sau:

**GIẢ ĐỊNH LIÊN TỤC**

*Các điểm gần nhau hơn có nhiều khả năng chia sẻ nhãn hơn.*Điều này cũng thường được giả định trong học tập có giám sát và mang lại ưu tiên cho các [ranh giới quyết định](https://en.wikipedia.org/wiki/Decision_boundary" \o "Ranh giới quyết định) hình học đơn giản . Trong trường hợp học tập bán giám sát, giả thiết về độ mịn bổ sung tạo ra ưu tiên cho các ranh giới quyết định ở các vùng có mật độ thấp, do đó có ít điểm gần nhau hơn nhưng ở các lớp khác nhau.

**GIẢ ĐỊNH CỤM**

*Dữ liệu có xu hướng hình thành các cụm rời rạc và các điểm trong cùng một cụm có nhiều khả năng chia sẻ nhãn hơn* (mặc dù chia sẻ dữ liệu nhãn có thể được trải rộng trên nhiều cụm). Đây là một trường hợp đặc biệt của giả thiết êm ái và làm tăng [tính năng học tập](https://en.wikipedia.org/wiki/Feature_learning" \o "Học tập tính năng) với các thuật toán phân cụm

**(THÊM) GIẢ ĐỊNH MANIFOLD**

3.Ứng dụng

\_Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ internet

Chia tách mật độ thấp

\_Một lớp chính của phương pháp cố gắng đặt ranh giới trong khu vực nơi có ít điểm dữ liệu (có nhãn hoặc không gắn nhãn). Một trong những thuật toán được sử dụng phổ biến nhất là [máy vectơ hỗ trợ chuyển tải](https://en.wikipedia.org/wiki/Support_vector_machine" \l "Transductive_support_vector_machines" \o ") , hoặc TSVM (mặc dù tên của nó, cũng có thể được sử dụng cho việc học quy nạp).

 \_Trong khi đó, [máy vector hỗ trợ](https://en.wikipedia.org/wiki/Support_vector_machines" \o "Hỗ trợ máy vectơ) cho việc học có giám sát tìm kiếm một ranh giới quyết định với tối đa [lợi nhuận](https://en.wikipedia.org/wiki/Margin_(machine_learning)" \o "Ký quỹ (học máy))so với dữ liệu được dán nhãn, mục tiêu của TSVM là một nhãn mác của các dữ liệu không có nhãn như vậy mà ranh giới quyết định có lợi nhuận tối đa trên tất cả các dữ liệu.

