**MACHINE LEARNING**

**Họ và tên:** ……*Đỗ Hoàng Long*……… **Nhóm:**….12......

**Lớp:** ……..*60TH4*………

**MSV:** ………1851061623…………..

**Phần 1: Tổng quan**

1. Giới thiệu về học máy

-Lịch sử và vai trò của machine learning

+ Machine Learning là một tập con của trí tuệ nhân tạo, nơi mà các thuật toán máy tính được sử dụng để tự học từ dữ liệu và thông tin. Trong machine learning, các máy tính không cần phải được lập trình một cách rõ ràng nhưng có thể tự thay đổi và cải thiện các thuật toán của chúng.

+ Hiện nay, các thuật toán machine learning cho phép máy tính có thể giao tiếp với con người, xe hơi tự lái, viết và xuất bản tường thuật các trận đấu thể thao, và tìm thấy kẻ tình nghi khủng bố. Tôi tin chắc rằng machine learning sẽ tác động một cách sâu sắc đến mọi ngành công nghiệp và các công việc liên quan đến chúng, đó là lý do tại sao mọi nhà quản lý cần phải có ít nhất một số kiến thức về machine learning và nó đã phát triển như thế nào.

+ 1950 - Nhà bác học Alan Turing đã tạo ra "Turing Test (phép thử Turing)" để xác định xem liệu một máy tính có trí thông minh thực sự hay không. Để vượt qua bài kiểm tra đó, một máy tính phải có khả năng đánh lừa một con người tin rằng nó cũng là con người.

+ 1952 - Arthur Samuel đã viết ra chương trình học máy (computer learning) đầu tiên. Chương trình này là trò chơi cờ đam, và hãng máy tính IBM đã cải tiến trò chơi này để nó có thể tự học và tổ chức những nước đi trong chiến lược để giành chiến thắng.

+ 1957 - Frank Rosenblatt đã thiết kế mạng nơron (neural network) đầu tiên cho máy tính, trong đó mô phỏng quá trình suy nghĩ của bộ não con người.

+ 1967 - Thuật toán "nearest neighbor" đã được viết, cho phép các máy tính bắt đầu sử dụng những mẫu nhận dạng (pattern recognition) rất cơ bản. Nó được sử dụng để vẽ ra lộ trình cho một người bán hàng có thể bắt đầu đi từ một thành phố ngẫu nhiên nhưng đảm bảo anh ta sẽ đi qua tất cả các thành phố khác theo một quãng đường ngắn nhất.

+ 1979 - Sinh viên tại trường đại học Stanford đã phát minh ra giỏ hàng "Stanford Cart" có thể điều hướng để tránh các chướng ngại vật trong một căn phòng.

+ 1981 - Gerald Dejong giới thiệu về khái niệm Explanation Based Learning (EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung để nó có thể làm theo bằng cách loại bỏ đi những dữ liệu không quan trọng.

+ 1985 - Terry Sejnowski đã phát minh ra NetTalk, nó có thể học cách phát âm các từ giống như cách một đứa trẻ tập nói.

+ 1990s - Machine Learning đã dịch chuyển từ cách tiếp cận hướng kiến thức (knowledge-driven) sang cách tiếp cận hướng dữ liệu (data-driven). Các nhà khoa học bắt đầu tạo ra các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra các kết luận - hay là "học" từ các kết quả đó.

+ 1997 - Deep Blue của hãng IBM đã đánh bại nhà vô địch cờ vua thế giới.

+ 2006 - Geoffrey Hinton đã đưa ra một thuật ngữ "deep learning" để giải thích các thuật toán mới cho phép máy tính "nhìn thấy" và phân biệt các đối tượng và văn bản trong các hình ảnh và video.

+ 2010 - Microsoft Kinect có thể theo dõi 20 hành vi của con người ở một tốc độ 30 lần mỗi giây, cho phép con người tương tác với máy tính thông qua các hành động và cử chỉ.

+ 2011 - Máy tính Watson của hãng IBM đã đánh bại các đối thủ là con người tại Jeopardy.

+ 2011 - Google Brain đã được phát triển, và mạng deep nơron (deep neural network) của nó có thể học để phát hiện và phân loại nhiều đối tượng theo cách mà một con mèo thực hiện.

+ 2012 - X Lab của Google phát triển một thuật toán machine learning có khả năng tự động duyệt qua các video trên YouTube để xác định xem video nào có chứa những con mèo.

+ 2014 - Facebook phát triển DeepFace, một phần mềm thuật toán có thể nhận dạng hoặc xác minh các cá nhân dựa vào hình ảnh ở mức độ giống như con người có thể.

+ 2015 - Amazon ra mắt nền tảng machine learning riêng của mình.

+ 2015 - Microsoft tạo ra Distributed Machine Learning Toolkit, trong đó cho phép phân phối hiệu quả các vấn đề machine learning trên nhiều máy tính.

+ 2015 - Hơn 3.000 nhà nghiên cứu AI và Robotics, được sự ủng hộ bởi những nhà khoa học nổi tiếng như Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác), đã ký vào một bức thư ngỏ để cảnh báo về sự nguy hiểm của vũ khí tự động trong việc lựa chọn và tham gia vào các mục tiêu mà không có sự can thiệp của con người.

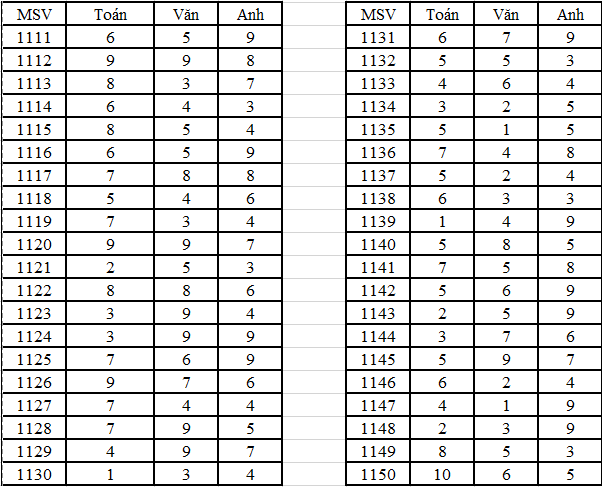
+ 2016 - Thuật toán trí tuệ nhân tạo của Google đã đánh bại nhà vô địch trò chơi Cờ Vây, được cho là trò chơi phức tạp nhất thế giới (khó hơn trò chơi cờ vua rất nhiều). Thuật toán AlphaGo được phát triển bởi Google DeepMind đã giành chiến thắng 4/5 trước nhà vô địch Cờ Vây.

-Ưu điểm và hạn chế của Học có giám sát và không giám sát

|  | Học có giám sát | Học không giám sát |
| --- | --- | --- |
| Định nghĩa | -Là hình thức tiếp nhận đầu vào và chuyển đổi dữ liệu đó thành đầu ra theo mong muốn người lập trình. Trong đó, có nhiều dạng đầu ra khác nhau như:  **+Phân loại:** thực hiện bài toán phân loại đối với các dữ liệu đầu vào khác nhau  **+Hồi quy:** thực hiện bài toán hồi quy cho các dải giá trị liên tục | -Là một phần thuật toán cơ bản của machine learning. Cũng đòi hỏi chuyển hóa dữ liệu đầu vào (không gán nhãn) thành dữ liệu đầu ra theo yêu cầu.  -Dạng này thường được sử dụng trong các công việc tìm hiểu, khám phá cấu trúc lẫn mối quan hệ của các dữ liệu khác nhau, phân cụm… |
| Ưu điểm | -Có thể dự đoán các giá trị đầu ra là các đại lượng liên tục như **doanh số hay giá cả** thay vì cố gắng phân loại chúng thành các đại lượng rời rạc như màu sắc và chất liệu của quần áo, hay xác định đối tượng trong một bức ảnh là mèo hay chó, … | -Trong thuật toán này, chúng ta không biết được *outcome* hay *nhãn* mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán. |
| Nhược điểm | -Chúng ta phải cung cấp cho nó rất nhiều những dữ liệu được gán nhãn (label).  -Việc thu thập dữ liệu gán nhãn mất nhiều thời gian, chi phí và công sức. Gây khó khăn không nhỏ cản trở machine learning. |  |

2.Trình bày bài toán

Input: Cho tập dữ liệu điểm thi của các sinh viên môn Toán , Văn và Tiếng Anh.



Ouput:

- Lập trình python theo phương pháp K-means phân thành 3 cụm và in ra màn hình 3 tâm cụm.

3. Trình bày phương pháp học máy được sư dụng trong bài tập lớn

-Mô tả phương pháp máy học mà bài tập lớn dung để dự đoán

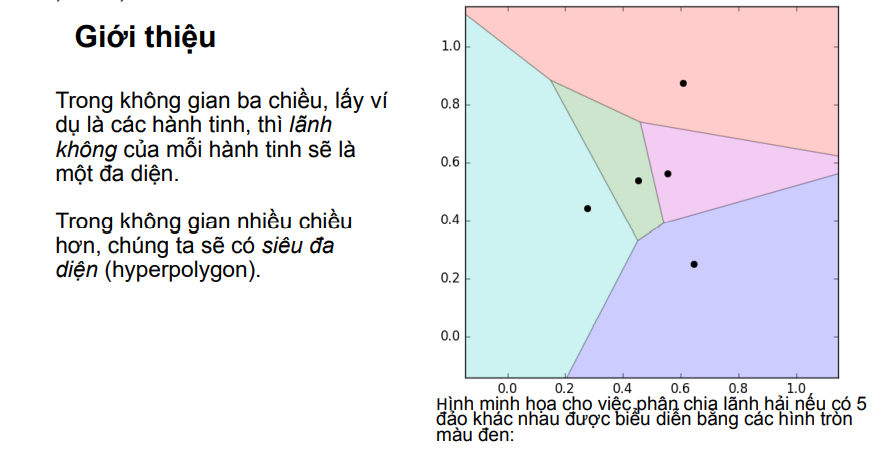
**Thuật toán K-MEANS:**

**A, Giới thiệu:**

• Thuật toán phân cụm K-means thuộc loại học không giám sát • không biết nhãn (label) của từng điểm dữ liệu • Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. • Ví dụ: Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau nhưng chưa có cách phân thành một số nhóm/cụm: • phân nhóm dựa trên sự tương tác giữa mỗi khách hàng với công ty: số năm là khách hàng; số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thành phố; nghề nghiệp; giả sử công ty có nhiều dữ liệu của rất nhiều khách hàng.

• phương pháp đơn giản là K-means Clustering sẽ được sử dụng để phân thành các cụm • Sau khi đã có các nhóm, nhân viên công ty lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào.

• Ý tưởng đơn giản nhất về cluster là tập hợp các điểm ở gần nhau trong một không gian nào đó.



Một số ký hiệu toán học trong K-Means

• Có 𝑁 điểm dữ liệu X = [x1, x2, … , x𝑁] ∈ ℝ𝑑×𝑁 và 𝐾 là số cụm cần phân chia.

• Chúng ta cần tìm tâm của mỗi cụm m1, m2, … , m𝐾 ∈ ℝ𝑑×1 và nhãn của mỗi điểm dữ liệu.

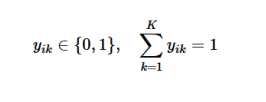
• Với mỗi x , đặt y𝑖 = [𝑦𝑖1, 𝑦𝑖2, … , 𝑦𝑖𝐾] là véc tơ nhãn

• Nếu x𝑖 được phân vào nhóm 𝑘 thì 𝑦𝑖𝑘 = 1 và 𝑦𝑖𝑗 = 0, ∀𝑗 ≠ 𝑘

• Ví dụ:

- Một điểm dữ liệu có véc tơ nhãn là [1,0, … , 0] thì điểm đó thuộc về lớp 1

• Ràng buộc của y𝑖 có thể viết dưới dạng toán học như sau:



**B, Hàm mất mát và bài toán tối ưu:**

• G/S m𝑘 là center (hoặc representative) của mỗi cluster

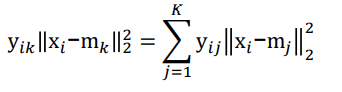
• ước lượng tất cả các điểm được phân vào cluster này bởi m𝑘

• một điểm dữ liệu x𝑖 được phân vào cluster k sẽ có sai số là (x𝑖−m𝑘)

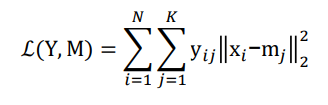
• Chúng ta mong muốn sai số (x𝑖−m𝑘) có trị tuyệt đối nhỏ nhất nên đại lượng dưới đây cần có giá trị nhỏ nhất:



• Vì x𝑖 được phân vào cụm k nên 𝑦𝑖𝑘 = 1, 𝑦𝑖𝑘 = 0, ∀𝑗 ≠ 𝑘, do đó chúng ta có:



• Sai số cho toàn bộ dữ liệu sẽ là:

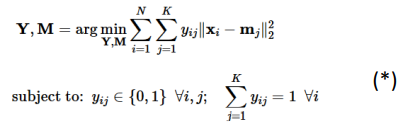


• Trong đó

• 𝑌 = [y1; y2; … ; y𝑁] là ma trận véc tơ nhãn

• M = [m1, m2, … , m𝐾] là ma trận tâm của mỗi cụm

• Chúng ta cần tối ưu bài toán để tìm Y và M



• Bài toán (\*) là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc.

• Bài toán này thuộc loại rất khó tìm nghiệm tối ưu toàn cục (global optimal point, tức nghiệm làm cho hàm mất mát đạt giá trị nhỏ nhất có thể).

**C, Thuật toán tối ưu hàm mất mát**

• Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu.

• Lưu ý: điểm cực tiểu chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất.

• Giải bài toán (\*) là:

• xen kẽ giải 𝑌 và 𝑀 khi biến còn lại được cố định

• đây là thuật toán lặp, cũng là kỹ thuật phổ biến khi giải bài toán tối ưu.

• Sau khi giải, chúng ta có được Y và M.

**D, Tóm tắt thuật toán**

**Đầu vào:** Dữ liệu X và số lượng cluster cần tìm K

**Đầu ra:** Các center M và label vector cho từng điểm dữ liệu Y

1. Chọn K điểm bất kỳ làm các center ban đầu

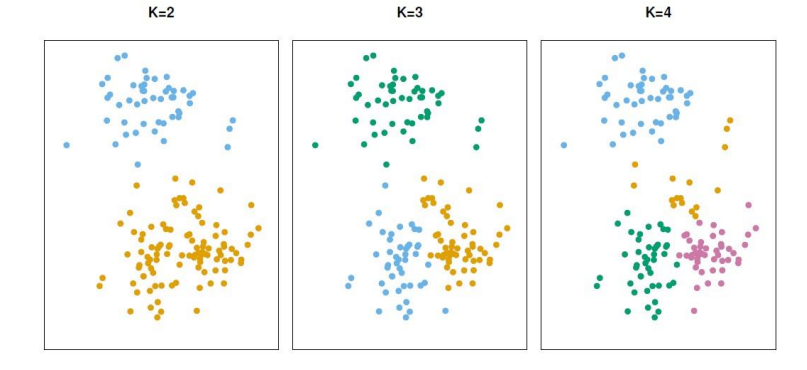
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất

3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì thuật toán dừng.

4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.

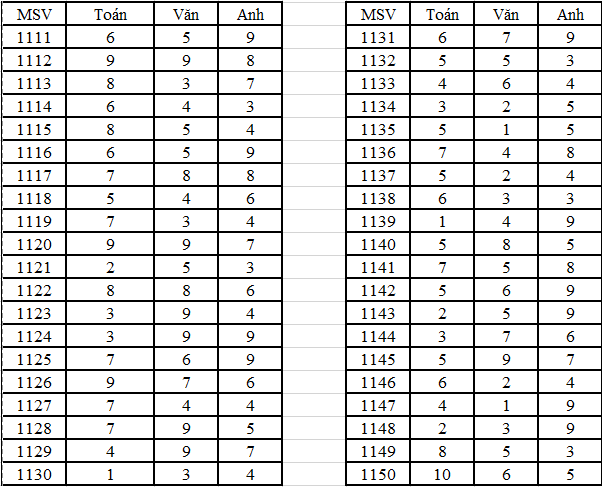
5. Quay lại bước 2.

**E, Minh họa**



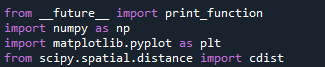
**Phần 2: Thực nghiệm**

1.Mô tả tập ví dụ huấn luyện và tập test

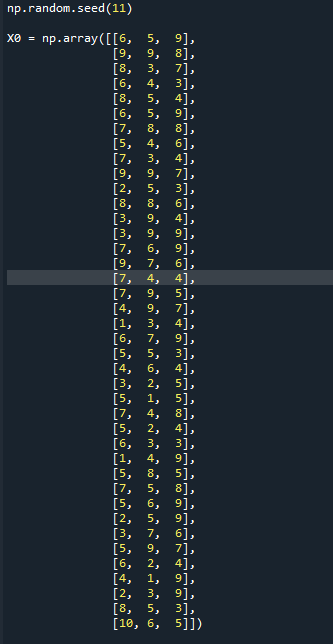


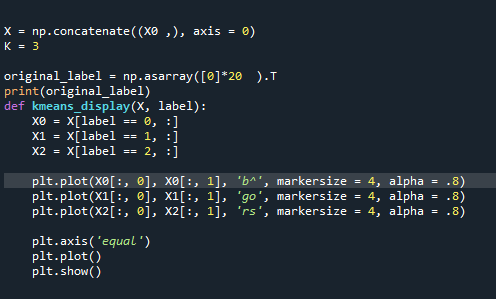
2.Mô tả chương trình demo

A, Import các thư viện:

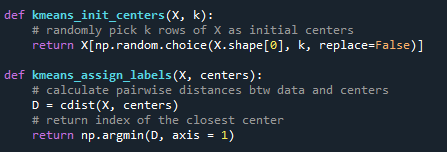


B, Nhập các dữ liệu đầu vào:

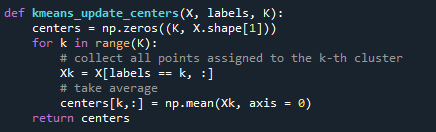




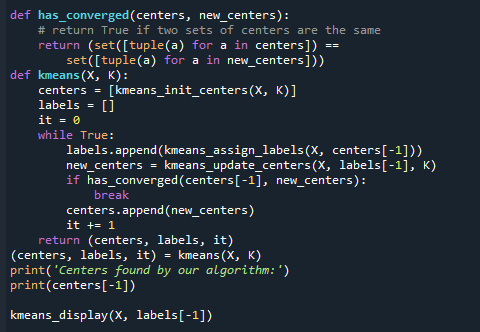
C, Chọn ngẫu nhiên k hàng X làm trung tâm ban đầu, sau đó tính toán khoảng cách theo cặp giữa dữ liệu và trung tâm, sau đó trả về chỉ số của trung tâm gần nhất.



D, Cập nhật tâm cụm mới

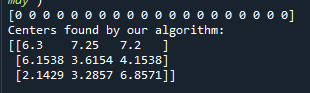


E, Kiểm tra xem tâm đã nằm giữa cụm hay chưa.

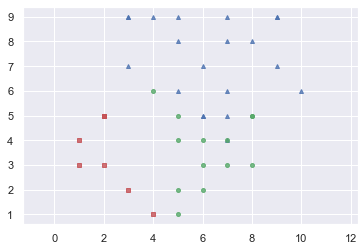


3. Mô tả kết quả dự đoán

A, Kết quả sau khi tìm ra tâm cụm:



B, Biểu đồ minh họa



**Kết luận**

**-Machine Learning** là một lĩnh vực mới và khá khó tiếp cận. Tuy nhiên các bạn hãy đi từ những thứ cơ bản nhất và tránh gặp phải những sai lầm trong quá trình học nó. Hi vọng bài viết này một phần nào đó giúp các bạn tránh được những hiểu nhầm đáng tiếc như trên.

**-** Vàsau 1 khoảng thời gian học về machine learning (học máy), em cảm thấy rất nhiều sự hữu ích của Machine Learning trong xã hội. Càng tìm hiểu, mình càng thấy nó có thể sử dụng để giải quyết mọi vấn đề.

- Sau cùng, em rất cảm ơn thầy giáo *Nguyễn Hữu Quỳnh* đã tốn nhiều thời gian và công sức để chỉnh dạy cho em rất nhiều điều hay và thú vị. Trong thời gian sắp tới, em sẽ tiếp tục tìm hiểu và học tập về Machine Learning (học máy)

**Tài liệu tham khảo**

* Slide bài giảng.
* <http://www.ungdung.khoa-hnvd.com/Hoc_thuat/KMeans.html>