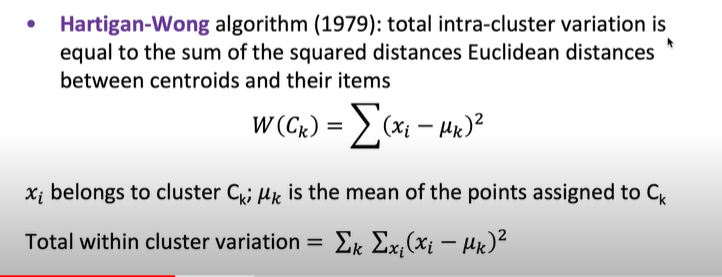
**Thuật toán phân cụm(clustering)**

**Thuật toán phân cụm K-Means là một trong những thuật toán phân cụm dữ liệu dựa trên học không giám sát được sử dụng nhiều trong các học máy nói chung và trong khai phá dữ liệu nói riêng.**

**Defining the clusters so that the total within-cluster variation is as minimum as possible**

**Hartiagan-Wong:**

****

Các bước:

B1: Đặt ra số lượng cluster  
B2: Khởi tạo ngẫu nhiên các giá trị trung tâm trùng 1 data nào đó

B3: Set các điểm gần nhất với gía trị khởi tạo thành 1 cluster

B4: Tính khoảng cách các giá trị của data đối với giá trị trung tâm

B5: Tính lại giá trị khởi tạo dựa trên trung bình giá trị khoảng cách các điểm với giá trị khởi tạo

B6: lặp lại

**Ví dụ: 7:40-13:10**

[**https://www.youtube.com/watch?v=B08yMWVGWpk**](https://www.youtube.com/watch?v=B08yMWVGWpk)

**Phân lớp (classification):**

**\*\*Học máy có giám sát :**

**(\*)** Học tập có giám sát là nơi bạn có các biến đầu vào (X) và biến đầu ra (Y) và bạn sử dụng thuật toán để tìm hiểu hàm ánh xạ từ đầu vào đến đầu ra.

Nó được gọi là việc học có giám sát bởi vì quá trình của thuật toán học từ tập dữ liệu đầu vào có thể được coi là một “giáo viên” giám sát quá trình học tập. Chúng ta biết câu trả lời đúng, thuật toán sẽ lặp đi lặp lại làm cho việc dự đoán về dữ liệu đầu vào liên tục được “giáo viên” hoàn thiện. Việc học dừng lại khi thuật toán đạt được mức hiệu suất ở mức chấp nhận được.

**Phân loại (Classification):** Việc phân loại diễn ra khi biến đầu ra là một thể loại nào đó, chẳng hạn như “đỏ” hoặc “xanh” hoặc “bệnh” và “không có bệnh”.

**Thuật toán Regression logistic**

Đây là một trong những thuật toán ML được sử dụng cho các vấn đề Classification khác nhau như phát hiện spam, dự đoán bệnh tiểu đường, phát hiện ung thư, v.v.

Regression logistic đơn giản hơn để triển khai, diễn giải và hiệu quả để huấn luyện.

Chúng tôi sử dụng Regression logistic để Classification nhị phân các điểm dữ liệu. Chúng tôi thực hiện Classification Classification sao cho đầu ra thuộc một trong hai lớp (1 hoặc 0).

Ví dụ – chúng ta có thể dự đoán liệu hôm nay trời có mưa hay không, dựa trên điều kiện thời tiết hiện tại.

<https://websitehcm.com/regression-logistic-kien-thuc-co-ban/>

### ****Thuật toán Bayes Naïve****

Thuật toán Naïve Bayes có thể là một thuật toán học có giám sát, được dự đoán trên định lý Bayes và được sử dụng để giải các bài toán Classification. Đó không phải là một thuật toán mà là một nhóm thuật toán trong đó tất cả chúng đều có chung một nguyên tắc tiêu chuẩn, tức là mọi cặp tính năng đang được Classification là độc lập với mọi tính năng khác.

Naïve Bayes Classifier là một thuật toán Classification đơn giản và tốt nhất giúp xây dựng các mô hình học máy nhanh sẽ đưa ra các dự đoán nhanh chóng.

Naive Bayes là một cách dễ dàng và nhanh chóng để dự đoán lớp của tập dữ liệu

Nhược điểm:

Nếu biến Classification thuộc danh mục không được theo dõi trong tập huấn luyện, thì mô hình sẽ cung cấp cho nó một xác suất bằng 0, điều này sẽ ngăn nó đưa ra bất kỳ dự đoán nào

Naive Bayes giả định sự độc lập giữa các tính năng của nó. Trong cuộc sống thực, rất khó để thu thập dữ liệu liên quan đến các tính năng hoàn toàn độc lập

<https://websitehcm.com/classification-cac-he-thong-data-mining/>

### ****Thuật toán K-Nearest Neighbors****

### K-nearest neighbor là một thuật toán supervised-learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training, mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới

Có một vài khái niệm tương ứng người-máy như sau:

| **Ngôn ngữ người** | **Ngôn ngữ Máy Học** | **in Machine Learning** |
| --- | --- | --- |
| Câu hỏi | Điểm dữ liệu | Data point |
| Đáp án | Đầu ra, nhãn | Output, Label |
| Ôn thi | Huấn luyện | Training |
| Tập tài liệu mang vào phòng thi | Tập dữ liệu tập huấn | Training set |
| Đề thi | Tập dữ liểu kiểm thử | Test set |
| Câu hỏi trong dề thi | Dữ liệu kiểm thử | Test data point |
| Câu hỏi có đáp án sai | Nhiễu | Noise, Outlier |
| Câu hỏi gần giống | Điểm dữ liệu gần nhất | Nearest Neighbor |

### Trong bài toán Classification, label của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set.

### 1 ví dụ dễ hiểu:

### <https://youtu.be/ETOqRZIrLY8>

### ****//****

### ****Support vector machine****

### ****Thuật toán Random Forest****

### ****Thuật toán Stochastic Gradient Descent****

### ****// Các thuật toán chưa phù hợp yêu cầu đề tài****

=> Rút kết luận chủ quan:

Những thuật toán này bản thân thấy phần lớn chưa cót ính thức tiễn trong triển khai thực tế, tuy nhiên có thể dùng **Hartigan-Wong** **clustering** để đánh giá khi bắt đầu tạo thang điểm và đánh giá định kỳ thang điểm sau này nếu như có quá nhiều thông số data bị thay đổi theo thời gian.

Thuật toán K-Nearest Neighbor(KNN) **có vẻ** phù hợp trong pha phân tích đánh giá điểm của mỗi option/provider khi đã có 1 bảng đánh giá thang điểm chung

\* Việc đánh giá định kì vẫn cần vẫn cần lập trình viên giám sát trực tiếp và modify thủ công vì crawler không có khả năng cralw hết các thông tin cần thiết cũng như việc các thông số không tồn tại ban đầu có thể phát sinh