**Nhóm:**

1. Trần Phương Ngọc Anh - 51702061
2. Nguyễn Kim Huệ - 51702108
3. Huỳnh Anh Tài - 51702171

**---------------------------------------------------**

**PLAN CHO BÀI TOÁN SENTIMENT**

1. **Mô tả bài toán**

Bài toán phân tích cảm xúc thuộc dạng bài toán phân tích ngữ nghĩa văn bản. Vì vậy, ta cần phải xây dựng một mô hình để hiểu được ý nghĩa của câu văn, đoạn văn để quyết định xem câu văn đó hoặc đoạn văn đó mang màu sắc cảm xúc chủ đạo nào.

Phát biểu theo góc nhìn của máy học (Machine Learning) thì phân tích cảm xúc là bài toán phân lớp cảm xúc dựa trên văn bản ngôn ngữ tự nhiên. Đầu vào của bài toán là một câu hay một đoạn văn bản, còn đầu ra là các giá trị xác suất của N lớp cảm xúc mà ta cần xác định. Trong loại bài toán phân tích cảm xúc này, chúng tôi sẽ phân tích cảm xúc (thái độ) trong văn bản thành 3 lớp: tích cực (positive), tiêu cực (negative) và trung tính (neutral).

1. **Xây dựng dữ liệu**

Bộ dữ liệu chính gồm 3880 sentiment, là sự kết hợp của 2 bộ dữ liệu sau:

* Bộ dữ liệu finegrained sentiment bao gồm các bài đánh giá sản phẩm từ nhiều nguồn trực tuyến khác nhau được Oscar Tackstrom và Ryan McDonald chú thích thủ công với sentiment ở nhiều cấp độ câu. Dữ liệu được thu thập từ nhiều thể loại khác nhau (sách, DVD, điện tử, âm nhạc, trò chơi điện tử) và được đánh giá tổng thể theo 3 lớp: tích cực, tiêu cực và trung tính.
* Bộ dữ liệu Sentiment Labelled Sentences UCI được xây dựng cho bài báo 'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015

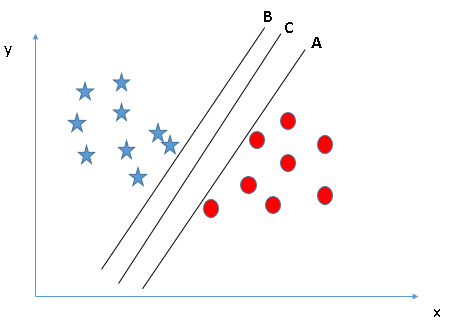
1. **Feature**

Extraction

Selection

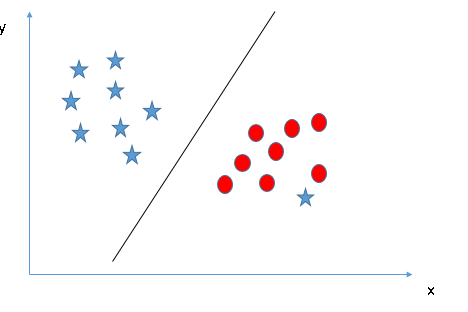
1. **Machine Learning Model**
2. ***SVM (Support Vectors Machine)***

SVM chỉ áp dụng cho thuật toán phân loại. Điểm mấu chốt của thuật toán này là làm sao vẽ cái đường vectors chia cắt các lớp một cách chính xác nhất.



Các quy tắc:

* Quy tắc 1: chia các lớp riêng biệt
* Quy tắc 2: Xác định khoảng cách giữa điểm gần vector nhất đến đường vector và loại bỏ các trường hợp ngoại lệ.



Ví dụ có 2 đường thẳng qua vector đó



Cách tìm khoảng cách (margin) đó

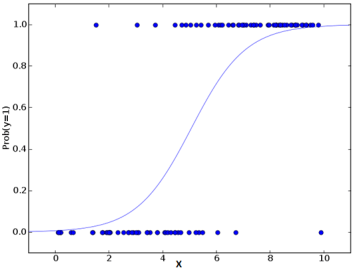
* Không gian 2 chiều: 
* Không gian nhiều chiều:

Cần phải tìm phương trình siêu phẳng trước 

Sau đó tính margin: 

1. ***Logistic Regression***

Là một thuật toán phân loại – classification, trong đó có 2 lớp class dưới dạng tuyến tính cách nhau, được sử dụng khi giá trị của biến mục tiêu có tính chất phân loại. Hồi quy logistic được sử dụng phổ biến nhất khi dữ liệu được đề cập có đầu ra nhị phân, do đó kết quả của biến mục tiêu là giá trị xác suất để phân loại đối tượng nghiên cứu. Các nhà toán học coi xác suất là các số trong khoảng [0,1], được gán tương ứng với một biến cố mà khả năng xảy ra hoặc không xảy ra là ngẫu nhiên. Kết quả của phương trình Logistic Regression sẽ dựa vào xác suất để quyết định giá trị sau cùng.



1. ***KNN (K-Nearest Neighbors)***

Thuật toán KNN dựa trên giả định là những thứ tương tự hay có tính chất gần giống nhau sẽ nằm ở vị trí gần nhau, như vậy KNN được xây dựng để tính toán khoảng cách giữa hai điểm dữ liệu để xem xét mức độ giống nhau của chúng, sau đó phân loại dựa trên những điểm tương đồng này.

Các bước trong KNN:

1. Ta có D là tập các điểm dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.
2. Đo khoảng cách (Euclidian, Manhattan, Minkowski, Minkowski hoặc Trọng số) từ dữ liệu mới A đến tất cả các dữ liệu khác đã được phân loại trong D.
3. Chọn K (K là tham số mà bạn định nghĩa) khoảng cách nhỏ nhất.
4. Kiểm tra danh sách các lớp có khoảng cách ngắn nhất và đếm số lượng của mỗi lớp xuất hiện.
5. Lấy đúng lớp (lớp xuất hiện nhiều lần nhất).
6. Lớp của dữ liệu mới là lớp mà bạn đã nhận được ở bước 5.

