Khi phân lớp phải tìm bộ dữ liệu huấn luyện

Tách thành 2 phần (Cột cuối cùng là cột nhãn label cho biết mỗi sample như vậy, mỗi đối tượng như vậy thuộc lớp nào)

Số lượng các lớp của tập dữ liệu huấn luyện lấy từ cột label

Ngoài cột nhãn ra, mình cần sưu tầm feautures bao nhiêu cột phía trước nữa (bao nhiêu thông tin mô tả object) ?

Họ tên, giới tính, cmnd, sdt, email, lớp học , nhiều quá, ta cần phân lớp những sinh viên đó theo kiểu gì , chứ ko phải lụm hết tất cả đặc

 trưng vô. Vì vậy số lượng đặc trưng cần chọn lọc một vài cái ta thực sự cần phân tích thôi

PCA: dimension reduction. Vì số lượng result càng lớn thuật toán càng chậm dễ dấn đến sai lệch (outlier)

dữ liệu nhỏ vừa thì làm gì, dữ liệu lớn thì làm gì ?

Khai phá dữ liệu sử dụng một số kĩ thuật học máy.

------------------------------------------------------------------

Neural Network , Deep Learning, Convolutional Neural Network

SVM học cái gì, học cái qui luật phân bố của các dữ liệu, học vị trí của các phân bố của các điểm dữ liệu trong không gian vector. Kết quả

là tìm ra siêu phẳng để tách dữ liệu thành 2 lớp.

Để tìm siêu phẳng, người ta  chỉnh bias, weight. SVM học sau đó dùng siêu phẳng nhận dữ liệu mới và nó sẽ dự đoán nằm bên + hay bên -

SVM không thể phân tách dữ liệu không có đường phân tách tuyến tính, không có bất cứ kẻ hở nào để kẻ đường phân cách giữa chúng.

Kernel chuyển từ  space này sang feauture space khác. (Tăng trục tọa độ lên thành không gian có số chiều lớn hơn ).

knn chỉ nhớ vị trí của các điểm dữ liệu mà thôi

còn phân lớp nó sẽ tính khoảng cách từ điểm mới đó đến k điểm gần nhất

Decision tree, từ 1 bảng dữ liệu huấn luyện

Bẻ bảng đó ra để tìm mối liên hệ mạnh yếu giữa cột này với các cột khác (Atromy để đo độ lợi thông tin, cao nhất, lấy làm gốc, theo từng layer)

Lấy nhãn gắn cho nó (nhãn nào nhiều hơn gắn vô).

---- Percepton -------

Từ 1 input, đo difference brror tới y output. Dùng cái sai số difference brror để điều chỉnh w sao cho input (x1,y1) gần y1,y label hơn)

Từ đó tìm được w đúng để còn dự đoán các x khác nhau. Người ta chỉ lấy 1 trong số nhiều vector xi. Sau đó lấy w1,w2,wn nhân với từng x11,x12,... x1n

Cộng hết từng giá trị tính được vào nhau bỏ vô hàm tạo ra yi recepton Yp

Tạo thành y có nhãn y label. Percepton: xi  --> yi'

từ y1 có sẵn kiểm tra với y được dự đoán xem nó có lệch so với yp hay ko?

vector w là strength, thứ hai là hàm activation function.

Bộ w1,..wn chọn đại nhân lại với nhau ra một con số kết quả nó là 1 cái lable reception mà nó đoán vector đầu vào (trong tập huấn luyện) có nhãn thực sự có gần giống với y của reception hay ko?

Reception chỉ có thể trong 1 khoảng.

Sigmoid: Những giá trị sau khi lấy w\*x có thể nằm trong khoảng đồ thị.

Activation nó ép toàn bộ giá trị thành w\*x không còn ngẫu nhiên (quá lớn hay quá bé nữa) vào 1 khoảng (0 đến 1 ) chẳng hạn

nếu ko có activation thì khoảng giá trị đầu ra quá lộn xộn.

Lấy từng giá trị z trong tổng xích ma, bỏ trong hàm sigmoid để tính nhãn, số càng nhỏ thì nhãn  = 0.

Do mình quyết định chọn hàm sigmoid hay tanh

Hàm tanh (tan) f(z)=tanh(z)

Mỗi format activation chỉ phù hợp với mỗi loại dữ liệu nhất định mà thôi

reLU (restricted linear unit), z là cái tổng nhan

if z <=0 relu = 0, nếu lớn hơn 0 thì nhãn là chính z luôn

Convolutional Neural Network chỉ dùng cho dữ liệu ảnh là tốt nhất

1 reception tính luôn cả giá trị đầu vô và 2 cái w luôn nhan

Yp - yi

Đạo hàm là tốc độ thay đổi dữ liệu tại một thời điểm thích hợp: Tính được sự thay đổi lim(f(x)-f)

Hàm đo độ hao hụt, gradient descent là giảm đạo hàm để mà giảm sai số cho bộ huấn luyện dữ liệu mà dùng w nào thì đồ thị giảm dần

SVM-SVM

## Bản chất toán học SVM ?

- Tham số ?

- SVM là một công cụ dùng để classification

+ SVM là một trong những công cụ sử dujg phương thức học có giám sát (Supervised) để thực hiện classiffication

+ có hai loại 1 là phân thành 2 lớp là binary classification

+ hai là nhiều lớp multi-classificcation

# phân lớp là phân lớp gì? tại sao?

- Mình có một tập dữ liệu, trong đó mình có cột loại, lớp từng sample (cột cuối cùng đặc trưng) ví dụ cột loại nghề nghiệp, loại sinh viên. Tập đó là tập dữ liệu huấn luyện để đưa vào SVm tạo model

- Huấn luyện tạo ra bộ classifier

- Huấn luyện cái công thức toán học của svm, những tham số của công thức đó sẽ được thay đổi, sẽ được học dựa theo bộ dữ liệu đó (data-driven)

- Nhưng tham số là gì, ví dụ y=ax+b, a, b là tham số. ví dụ y-ã+b gọi là bộ phân loại.

- Những dữ liệu không có trong bộ dataset training khi được đưa vào bộ phân loại ấy. Nó sẽ tự gán nhãn cho phù hợp, hoặc không gán nhãn nêu sample đó.

- Cho một tập dữ liệu gồm có những vector i chạy từ 1 đến n

- Sử dunjgxasc suất thống kê, có vùng phân bổ, để nếu sample đó thuộc trong tập phân bố đó thì mới gán nhãn

- SVM không phải là thuật toán duy nhất , có nhiều biến thể khác dựa trên svm được phát triển thêm

# SVM

- Biến đổi những thông tin chữ, số hóa thành vector

- chữ x nét đôi, in đậm, mcos mũi tên thì

- Trong bộ dữ liệu ,không phải lúc nào cũng lấy tất cả các feautures của bộ dữ liệu

- Số hóa từng feautures của nó

- Tại sao n từ 1 đến vô cùng

- n trên R là số chiều (dimensity) của vector

- dữ liệu +, - được lấy từ tập huấn luyện

- Cần tách dữ liệu trời nắng, trời mưa ra

- Mã hóa trời nắng băng dấu +

- Mã hóa trời mua bằng dấu -

- Mã hóa từng dữ liệu (-1 hoặc +1 )

- Chỉ có 2 lớp nắng mưa thì chỉ có -1 và +1

- Mỗi sample được gán nhãn: (vector x, y) vetor x có nhãn y

- n nhỏ là số chiều

- N là số lượng sample

- xi là dữ liệu ở hàng đầu tiên được mã hóa, yi là dữ liệu trong cái hàng đó

- Training dataset là input data của svm

- Biến thành bảng dữ liệu ở dạng vector

- Tìm đường phân cách để tách ra dấu +, dấu - (Đấy là nhiệm vụ cuối cùng của svm)

- Bảng dữ liệu vector đưa vào không gian vector (data point in vector space)

- R2: LINE R1 (y = ax + b)

- R3 PLANE R2 (z = ax + by + c)

- dữ liệu đầu vào là không gian n chiều (hyperlance: siêu phẳng) Rn ...... được biểu diễn ở không gian con R(n-1)->sub space

- Siêu phẳng là một cái mặt có hình dáng gì đó để phân cách dữ liệu thành 2 hay nhiều chiều

- Cài đặt siêu phẳng dưới dạng phương trình f(x) = w(t)x + b

- (vector xi, yi (là nhãn nhan))

vis duj y=ax+ bL thay đổi a là xoay cái đường thẳng

- thay đổi b là dịch chuyển lên xuống

f(vector (x)) = vector t \* vector x + vector b

vector (x) là đại diện cho tất cả vector xi trong không gian (i-1...N)

f(vector (x)) như là một cụ để tìm được mặt nào là cái mặt phân cách dữ liệu đó thành 2 mặt riêng biệt hoặc nhiều mặt tùy theo chiều của dữ liệu (hyperlance)

vector w là normal vector , là nghiêng mặt phẳng (vì vector w luôn vuông góc với mặt phẳng), gọi là weight vector (vector trọng số)

chỉnh bias là kéo lên xuống

vậy SVM nó tìm hai vector w và vector b

## AI bao trùm (Machine Learning, Data Mining)

\* Machine Learning: (Alogrithm)

\* Data Mining: Draw (Rule,Knowledd) \*\*from BigData\*\*

## Các bước trong Khai phá dữ liệu

\* Collection Data -> Prepare Data (PCA) -> Data Mining (Classification) -> Model -> Model Prediction

# Phân lớp -> Tạo model

- Khác phân cụm

- Bước chuẩn bị data dữ liệu trước đó là rất quan trọng

- Là học đi học lại 1 bộ dữ liệu có sẵn (đã xảy ra trong quá khứ)

- Data Driven = Learn from Data

- Cần dữ liệu chất lượng cao

# Lableling Dataset

- Feautures đầu tiên hoặc cuối cùng thường là phân loại theo lớp hoặc loại của dữ liệu ấy (Ví dụ loại nghề nghiệp)

- Ví dụ Lableing nam, nữ cho từng tấm hình (Dữ liệu dạng hình ảnh)

- Training dataset là dataset chỉ dùng cho huấn luyện , nó có cột label ở cuối cùng

- Dataset bình thường có thể không có cột label

# Model:

- Là kết quả sau khi huấn luyện được

- Có tên là bộ phân lớp (Classifier)

- Kết quả của phân lớp là kết quả của quá trình huấn luyện từ công thức toán ban đầu (công thức nó sẽ được uốn nắn theo phương trình, công thức toán nào đó phụ thuộc vào dữ liệu )

- Ví dụ để nhận dạng ngôi nhà nó sẽ trích xuất thông tin từ các cạnh (edge) (phải lấy dữ liệu là hình rất nhiều ngôi nhà khác nhau) để model này càng ngày càng nhận biết cạnh của ngôi nhà như là hình dáng của ngôi nhà

\*\*Lưu ý cần 1 tập huyến luyện, mỗi 1 sample cần gán nhãn cụ thể\*\*

\*\*Ngoài training dataset chọn 1 phương pháp phân lớp (giải thuật nào đó có công thức)\*\*

\*\*Công thức đó sẽ được điều chỉnh tham số trong quá trình huấn luyện tạo ra 1 công thức cuối cùng là mô hình phân lớp (Model)\*\*

\*Model là 1 công thức toán, mà công thức toán đó được điều chỉnh thông qua quá trình huấn luyện, học xong. Kết quả cuối cùng nó tạo ra 1 tập model, dự đoán sample mới (sample mà nó không có trong tập huấn luyện nhan) thuộc lớp nào\*

\*\*Chụp hình, gán nhãn cho các hình có cùng ý nghĩa (ví dụ chữ cái), tạo ra 26 chữ cái tương đương 26 lớp\*\*

\*Đưa 1 chữ a mới chưa có trong tập huấn luyện vào, nó sẽ tự hiểu 80% chữ a mới đó thuộc lớp a\*

\*Mỗi 1 giải thuật phân lớp có 1 công thức toán học\*

\*Classification Alogritham\* là mô hình phân lớp (Model) có tác dụng là nhận biết dữ liệu sample mới không thuộc

\*Class là label, 1 sample thuộc 1 lable/class nào đó\*

\*Classifier Model là Mode, là 1 công thức toán học sau một quá trình điều chỉnh tham số rất là lâu (Có tên là training proccess)\*

\*Classification là phân lớp: là các thuật toán được học trên bộ dữ liệu huấn luyện tạo Model, mà Model đó dùng để dự đoán các dữ liệu mới mà nó chưa thấy bao giờ (), dự đoán ở đây là dự đoán cái nhãn của dữ liệu đó\*

- Phân chia tuyến tính là kẻ đường thẳng chia làm 2 khu vực để phân lớp (Dữ liệu có thể phân lớp 1 cách tuyến tính)

- Phương trình tuyến tính là 1 đường thẳng: ví dụ vận tốc càng \*tăng\* thì thời gian đi càng \*giảm\*

- Trên là binary Classification (Phân thành 2 lớp 0 và 1)

- Phân những lớp khác nhau bằng nhiều đường thẳng (cũng xêm tuyến tính) gọi là Multi classification

{

(Level 1) : Dùng / Sử dụng

(Level 2) : Dùng, nhưng qua có tùy chỉnh chỉnh sửa nó, cần học cách làm ra \*nó\* ví dụ nó là \*quần áo\*

mà cái mà mình học là nó có sẵn rồi nhan (Existing) (Thạc sĩ)

(Level 3) : Đề xuất cách làm mới (new proposals) tương đương công trình nghiên cứu khoa học (Tiến sĩ)

}

##### SVM ####

Siêu phẳng là gì?

1. Là 1 trong những giải thuật phân lớp, phân 1 lớp, 2 lớp hoặc nhiều lớp

2. Là 1 giải thuật học máy dựa trên những vector hỗ trợ

3. Đi tìm phương trình vẽ ra đuòng w thẳng nó nằm gần những vector hỗ trợ

4. Vector là gì? là một chấm trong không gian 1 chiều, 2 chiều, hoặc nhiều chiều

chỉ là một chấm nó là vô hướng

5. Đường thẳng kẻ từ origin đến điểm đó là một vector có chiều

6. Supervised learning method là học dựa trên bộ dữ liệu huấn luyện mà cái nhãn của dữ liệu sample đó nó có sẵn.

7. Unsupervised learning method là học dựa trên bộ dữ liệu huấn luyện mà không có nhãn