Topics to cover.

1) Gaussian function.
2) Baye's Theorem.
3) Naive Bayes classifier.
4) Non-Linearity
5) Logistic Regression

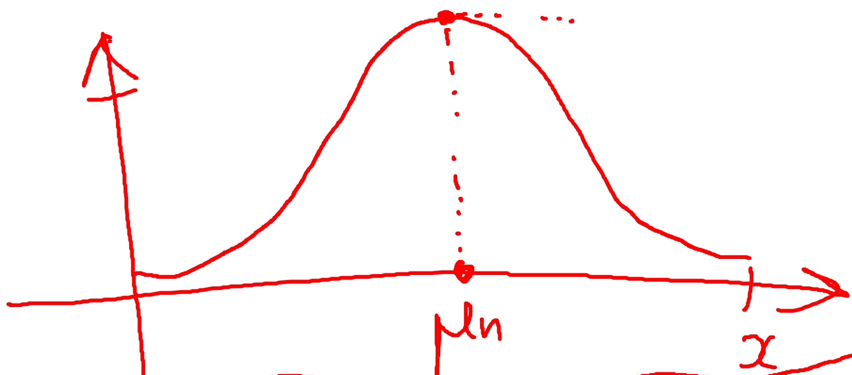
Gaussian Function.

1) Classification :-

$\boxed{\text{Inputs} = \text{Features}}$

$\boxed{\text{Outputs} = \text{Labels}}$.

Gaussians.



$$\boxed{f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Relative Spread :-

if a r.v. is extracted from a sample that follows this curve, there is high probability that it belongs to the "<u>middle-band</u>".

If we claim that the input data follows a Gaussian curve what parameters do we need to define the data?

1) $(x_1 y_1), (x_2 y_2), \ldots\ldots (x_n y_n) \longrightarrow$ data points.

2) μ, σ. values.

Now, how to find the μ and σ ??

For any candidate parameter, we can associate Likelihood. function $\rightarrow$ "support" provided by the input data for the given parameter.

Likelihood function is a Joint PMF/PDF.

$\swarrow$

So to find the "right" parameter, it needs to be the maxima of the Likelihood function. WHAT TO DO??

MLE (Max. likelihood Estimation).

Let us say uni-variate Gaussian

$y = \{y_1, y_2, \ldots, y_n\}$.

then, the maximum likelihood estimates for μ and σ^2 are....

$$\boxed{\mu = \frac{1}{n} \sum_{i=1}^{n} y_i} \quad \boxed{\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2}$$

$\bar{y} = \mu$.

Naive Bayes classifier.

- Let's say input vector is x.
- Set of K classes $C_1\ C_2 \ldots\ C_K$ are there.
- We need to find K where $P(C_K | x)$ is maximum.

(class K has highest p. of accomodating x).

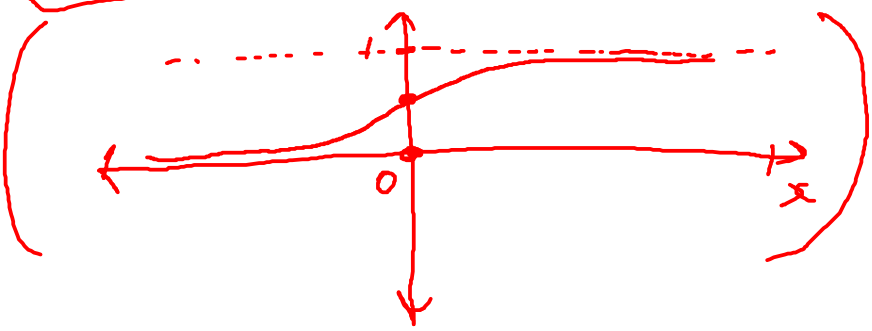
- Let's assume all features are independent of each other.
- Let's assume that each feature follows a Gaussian.

$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \cdot P(\text{class})}{P(\text{data})}$$

Logistic Regression

$$\boxed{f(x) = \frac{1}{1 + e^{-x}}}$$

Sigmoid



- Just like Linear Regression; we get a line.
- Now, what we do is we feed this line into the Sigmoid function to get a value between 0 and 1.

Then we threshold the value to particular class (0/1).

$$y = b_0 + b_1 x \quad \longleftarrow \underline{\text{Linear Model}}$$

$$\left. P = \frac{1}{1 + e^{-(b_0 + b_1 x)}} \right\} \underline{\text{Logistic Model}}$$

Q. How is this approach different from the Naive Bayes Model?

→ We are directly calculating $p(y|x)$

---

* Logistic Regression (Mustafa)

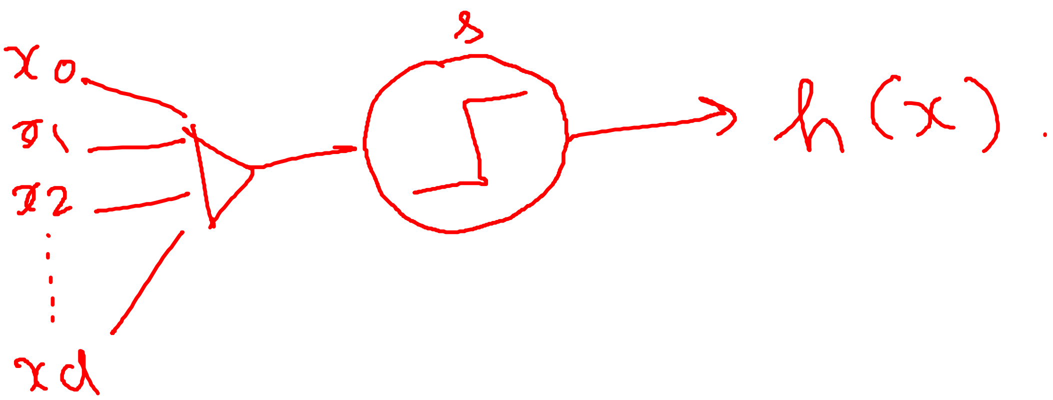
1) Logistic Idea + Linear Regression.
2) Discriminative / Generative Model.
3) Parameters - feature weights.
4) Estimation - Max. likelihood Esti...

↓

Linear Model $s = \sum_{i=0}^{d} w_i x_i$

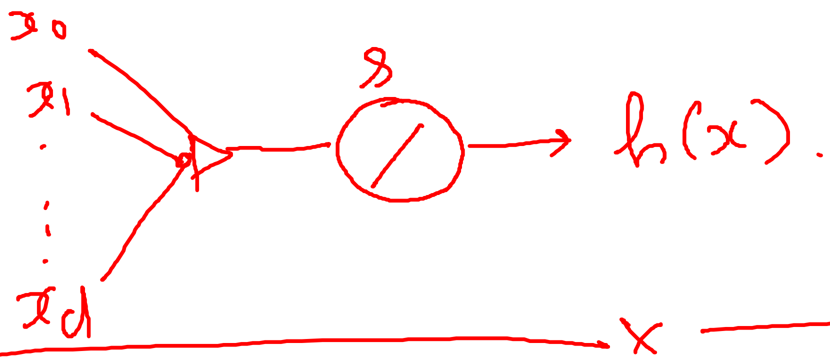
1) Linear Classification.

$h(x) = \text{sign}(s)$.



2) Linear Regression.

$h(x) = s$.

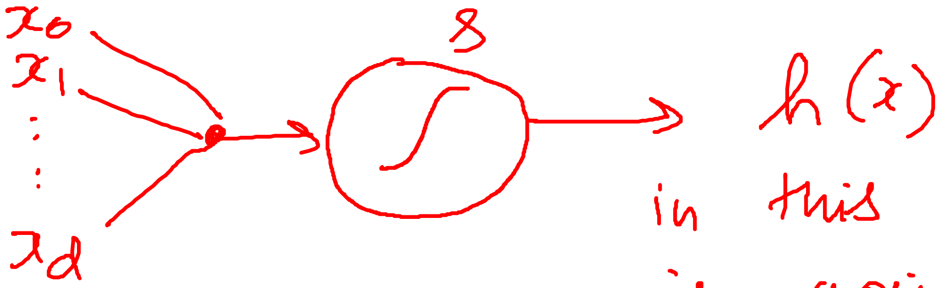


3) Logistic Regression

$h(x) = \theta(s)$.

We will take s and apply non-linearity to it.

$h(x) = \theta(s)$.



in this case, output is going to be <u>interpreted as a probability</u>.

$$\theta(s) = \frac{1}{1+e^{-s}} = \frac{1}{1+\frac{1}{e^{s}}}$$

$$\boxed{\theta(s) = \frac{e^{s}}{1+e^{s}}}.$$

- Hard-threshold will be decision making $\rightarrow$ <u>either 0 or 1</u>.

- The sigmoid function brings in the <u>soft threshold (uncertainty)</u>

- Why is it important?

Sometimes Probability is more important than the direct decision 0/1.

- Let's say the problem is about what is the probability that a person will get a heartattack....

$$\boxed{s = w^T x}$$

$[1 \times n][n \times 1]$
$\downarrow$
$[1 \times 1]$

Let's say (x, y) is the data we have. (Then y is binary)

$y = \pm 1$.

$$P(y|x) = \begin{cases} f(x) & \text{for } y = +1. \\ 1 - f(x) & \text{for } y = -1. \end{cases}$$

*. We are trying to learn the $f(\,)$ here...

$f : \mathbb{R}^d \longrightarrow [0,1]$ is the probability.

Learn $g(x) = \theta(w^T x) \approx f(x)$.

Error Measure (Loss function)

for each (x, y) $y = \pm 1$.

y is generated by probability $f(x)$.

We have a plausible error measure based on likelihood.

If $h = f$.

What is the probability of generating this data if your assumption is true.

If that probability is small, the assumption is pure.

(Vice Versa).

if $h = f$, how likely is it to get y from x?

If one chooses to uses the probabilistic approach for choosing the hypothesis;

"What is the most probable hypothesis given the data".

Here, we are asking ...

"What is the probability of the data given the hypothesis".

$\boxed{\text{WHICH IS BACKWARDS}}$

$$P(y|x) = \begin{cases} f(x) & \text{for } y = +1, \\ 1 - f(x) & \text{for } y = -1. \end{cases}$$

$$P(y|x) = \begin{cases} h(x) & \text{for } y = +1 \\ 1 - h(x) & \text{for } y = -1. \end{cases}$$

Formula for Likelihood...

$$h(x) = \theta(W^T x).$$

Note:- $\theta(-s) = 1 - \theta(s)$.

$$f(x) = \frac{1}{1+e^{-x}}.$$

$$f(-x) = \frac{1}{1+e^{x}} = \frac{1}{1+\frac{1}{e^{-x}}},$$

$$= \frac{e^{-x}}{1+e^{-x}}.$$

$$= \frac{+1 - 1 + e^{-x}}{1+e^{-x}}.$$

$$= \left(\frac{1+e^{-x}}{1+e^{-x}}\right) - \frac{1}{1+e^{-x}},$$

$$f(-x) = 1 - f(x)$$

$$P(y|x) = \sigma(y W^T x).$$

Entire dataset

$(x_1 y_1)\ (x_2 y_2) \ldots \ldots \ldots (x_n y_n)$

$$\prod_{n=1}^{N} \sigma(y_n W^T x_n).$$

<u>Maximizing the Likelihood.</u>

$$\rightarrow \prod_{n=1}^{N} \sigma(y_n W^T x_n).$$

$$\rightarrow \boxed{\sum_{n=1}^{N} \ln(\sigma(y_n W^T x_n))}$$

This we have to maximize.

$$-\frac{1}{N} \cdot \Sigma \ln\left(\frac{1}{\sigma(y_n W^T x_n)}\right)$$ <u>minimise</u>

$$\sigma(s) = \frac{e^s}{1+e^s} \Rightarrow \frac{e^s/e^s}{\frac{1+e^s}{e^s}} \Rightarrow \frac{1}{1+e^{-s}}.$$

$$\underbrace{\text{Error}}_{\text{this we have to minimize}} = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ln\left(1+e^{-y_n w^T x_n}\right)}_{e(h(x_n),\, y(n))}$$

this we have to minimize

$e(h(x_n), y(n))$.

cross-entropy error

WHAT TO DO ??

→ Gradient Descent !!

Now, what if it is more than 2 classes ?? Now, something called as Softmax Regression comes into the picture . . .

Logistic Regression

* 1 more attempt

$$cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1. \\ -\log(1-h_\theta(x)) & y=0. \end{cases}$$

$$cost(h_\theta(x), y)$$

$$= -y.\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$$

$$h_\theta(x) = \frac{1}{(1+e^{-\theta^T x})}.$$

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

$$\frac{d(\sigma(x))}{dx} = \frac{0(\ ) - (1).[e^{-x}(-1)]}{(1+e^{-x})^2}.$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1+e^{-x}-1}{(1+e^{-x})^2} = \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(\ )^2}$$

$$= \frac{1}{(1+e^{-x})} - \frac{1}{(1+e^{-x})^2}$$

$$= \left(\frac{1}{1+e^{-x}}\right)\left[1 - \frac{1}{(1+e^{-x})}\right]$$

$$\boxed{\frac{d(\sigma(x))}{dx} = \sigma(x).(1-\sigma(x)).}$$

Chain Rule

$$\frac{\partial(J(\theta))}{\partial(\theta_j)} = \frac{1}{m}. \sum_{i=1}^{m}\left[y_i\left[\frac{1}{h_\theta(x^i)}\right]\left[\frac{\partial(h_\theta(x^i))}{\partial\theta_j}\right]\right.$$

$$\left. + \sum_{i=1}^{m}(1-y_i)\left[\frac{1}{[1-h_\theta(x^i)]} \times \frac{\partial(1-h_\theta(x^i))}{\partial(\theta_j)}\right]\right]$$

$$= \downarrow$$

$$\frac{\partial(J(\theta))}{\partial(\theta_j)}$$

$$\Rightarrow \frac{-1}{m} * \left(\sum_{i=1}^{m} \left[y^{(i)} \times \frac{1}{h_\theta(x^i)} \times \sigma(z).[1-\sigma(z)] \frac{\partial(\theta^T x)}{\partial \theta_j} \right] \right.$$

$$\left. + \sum_{i=1}^{m} \left[(1-y^i) * \frac{1}{[1-h_\theta(x^i)]} (-\sigma(z).[1-\sigma(z)]).\frac{\partial(\theta^T x)}{\partial \theta_j} \right] \right)$$

$$\Rightarrow \frac{\partial(J(\theta))}{\partial(\theta_j)} = \frac{-1}{m} \times \left(\sum_{i=1}^{m} \left[y^i \times (1-h_\theta(x^i))(x_j^i) - (1-y^i)\, h_\theta(x^i)\, (x_j^i) \right] \right.$$

Horrible
difficult to
understand

$$\frac{\partial(J(\theta))}{\partial(\theta)} = \frac{1}{m} . X^T [h_\theta(x) - y]$$

[This is what we finally get !!]

$$\boxed{\frac{\partial(J(\theta))}{\partial(\theta)} = \frac{1}{m} X^T[h_\theta(x) - y]}$$

$$[X] = [\text{d\_train\_points} \times \text{Features}]$$

$$[W] = [4 \times 1]$$

$$\boxed{[W] = [W] - \alpha[dw]}$$

$$[dw] = [4 \times 1]$$

$$[A] = [\text{d\_trainpoint} \times 1]$$

$$A = [\underbrace{P_1 \quad P_2 \quad \ldots \quad P_N}_{\text{dtrain}}]$$

$$\text{temp} = [\underbrace{\text{d\_train points}}]$$

$$[X^T \times (h(\theta) - y)]$$

$$\downarrow$$

$$[[\text{Features} \times \text{dtrain}][\text{dtrain} \times 1]]$$

$$\left[\text{Features} \times 1\right]$$

$$\begin{bmatrix}\text{New Features}\\ \\ \end{bmatrix} = \begin{bmatrix}\text{old}\\ \\ \end{bmatrix} - \alpha\begin{bmatrix}\frac{\partial(J(\theta))}{\partial\theta}\end{bmatrix}$$

Naive Bay's

1) Assumption :- Gaussian Distribution of all features. And all features are considered independent.

$$P(y=C_k|x) = \frac{P(x|y=C_k)P(y=C_k)}{P(x)}$$

This let's say 2 classes.

0	1
4 features	4 features
4 Gaussians	4 Gaussians

Product for given datapoint.

$[x_1 \; x_2 \; x_3 \; x_4]$

$(CT_1 \; CT_2 \; CT_3 \; CT_4)$ $(CT_1 \; CT_2 \; CT_3 \; CT_4)$

$(\text{Prior } P)^4$ $(\text{Prior})^4$.

Whichever is greater

data belongs to that class.

X