① Supervised Learning

So given X ⟹ predict y ⟹ input $\xrightarrow{\text{model}}$ output
(label)

Regression
↳ output will be in a continuous Range

Classificatⁿ
↳ output will be in classes (discrete)

② Unsupervised learning

So, given X ⟶ do sth with it
(No label)

So, technically we only detect patterns!

↳ kMeans → k++ means

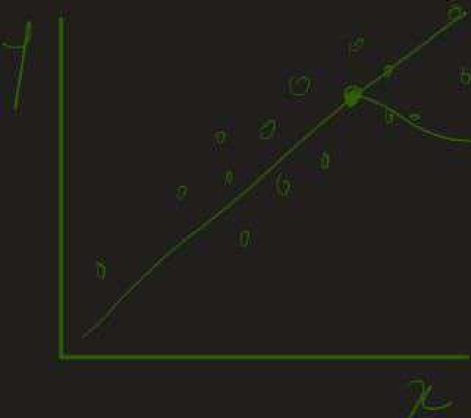
MSE Loss :- (1) Normal Distribution

$$N(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \rightarrow \text{Shape}$$

so $\sigma\uparrow \Rightarrow$ flattens

$\sigma\downarrow \rightarrow$ ⋀ $\Rightarrow$ more deterministic

for linear Regression $\Rightarrow$ Assumpt^n $\rightarrow$



$\rightarrow$ our samples are normally distributed @ mean $= \underline{w^T x}$!

$\Rightarrow$ to calulate the optimal value we do MLE :-
$\downarrow$
Maximum Likelyhood estimation

so, if we have a distribution & we want to calulate its parameter, as we only know distributi^n but don't know param

$\Rightarrow$ histogram

if we check the Quantity

$$M(\theta, y) = \prod P(y|\theta)$$

$\downarrow$

$M(2) > M(1),\ M(2) > M(3)$

bcause probab is more where points are more so $\prod$ will have more value!! so, thats why we define

$$\text{MLE} \Rightarrow \boxed{\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, M(\theta, y)}$$

Regression and clasification

14 December 2025 13:40

MLE for Linear regression ⇒

$$\hat{w} = \underset{w}{\arg\max} \prod_{1}^{n} P(y_i / w^T x_i)$$

$$\Rightarrow \underset{w}{\arg\max} \prod_{1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\left(\frac{y_i - w^T x_i}{\sigma_i}\right)^2} \quad ; \text{ assume } \sigma_i = \text{Const}$$

$$\underset{w}{\arg\max}\, l = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{1}^{n}\left(\frac{y_i - w^T x_i}{\sigma_i}\right)^2}$$

$$\underset{w}{\arg\max} \log L = -\frac{1}{\sigma_i^2} \sum_{1}^{n} \left(\frac{y_i - w^T x_i}{\sigma_i}\right)^2 = -\underbrace{N}_{\text{Const}}(MSE) + \text{Const}$$

So $\boxed{\underset{w}{\arg\max} \log L = \underset{w}{\arg\min}\ MSE}$

MLE for Classificatⁿ

⇒ binary model ⇒ $P_i \rightarrow$ probability of $y \in 0$ $(1-P_i) \Rightarrow \underline{y \in 1}$

So $l = P_i^m (1-P_i)^n$; $m \rightarrow$ # samples $\in 0$
$n \rightarrow$ # samples $\in 1$

$$\boxed{\underset{w}{\arg\max} \log l = m \log P_i + n \log(1 - P_i)}$$

Regression and clasification

14 December 2025 13:40

quick calculatⁿ

$$w^T x = [w_1, w_2 \dots w_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\nabla loss = \nabla\left(\frac{1}{N}\sum \left(y_i - w^T x_i\right)^2\right) = \nabla l$$

$$w^T x_i = (w_1 x_1 + w_2 x_2 \dots w_n x_n)$$

So

$$\frac{\partial l}{\partial w_i} = -\frac{2}{N}\left(\sum (y_i - w^T x_i) x_i\right)$$

if is as

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ x_1^2 & x_2^2 & \dots & x_m^2 \\ & \vdots & & \\ x_1^n & x_2^n & \dots & x_m^n \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \quad y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix}$$

then

$$\boxed{\begin{array}{l} ① \ MSE = \frac{1}{N}(y - Xw)^T (y - Xw) \\ ② \ \nabla l = \frac{2}{N}(X^T)[y - Xw] \end{array}}$$

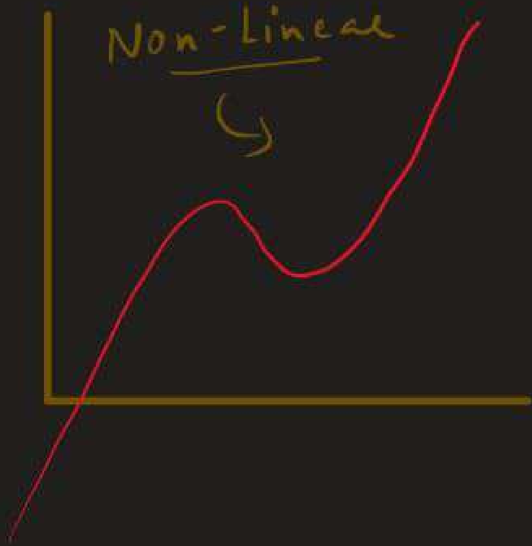
Is linear regression really linear?

14 December 2025 13:40

$\hat{y} = w^T x$

→ lets look at a polynomial f^n

$y = ax^3 + bx^2 + cx + d$ →



but if we map features

$\vec{x} = [1, x, x^2, x^3]$

then,

$$y = [d, c, b, a]\begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}$$ ⇒ so becomes liner!!

so $w^T x$ → hyperplane in feature dim ⇒ but if features are correlated ⇒ can become non-liner

Q :- $\hat{y} = w^T x$ & x is a polynomial of degree 'n' if training dataset contains 'm' points where

$m < n$ ⇒ what is the min loss possible

{ $y = w^T x$ } ??

Is linear regression really linear?

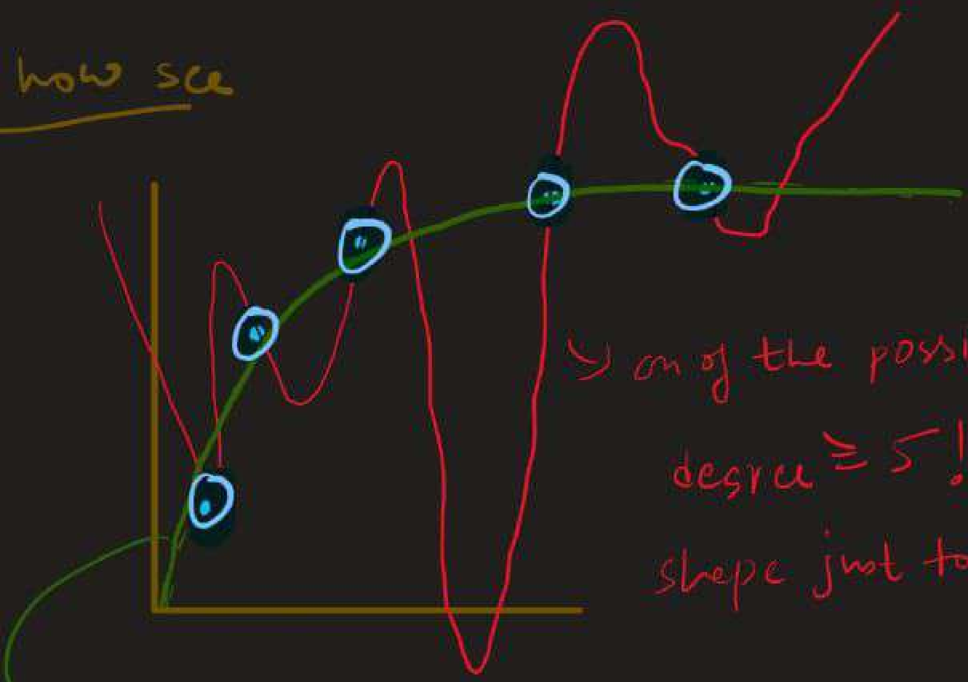
14 December 2025 13:40

for the problem since $\exists$ $m<n$ points $\Rightarrow (y - w^T x)$ have 'm' predefined roots now choose $n-m$ roots $= 0$ ie

$p(x) = (x-\alpha_1)(x-\alpha_2) \cdots (x-\alpha_m)\, x^{n-m} \rightarrow n$ degree pol

$= a_n x^n + a_{n-1} x^{n-1} \cdots a_1 x + a_0 \Rightarrow \boxed{MSE = 0}$!!

now see



one of the possible polynomials of degree ≥ 5! $\Rightarrow$ giving random shape just to set $MSE = 0$

while for this $MSE \neq 0$ but it represent the correct trend!

in the red case its observed that w_i become much large $\simeq 10^5 \rightarrow$ soo we make $\boxed{\text{newloss} = MSE + \sum |w_i|^2}$

this $\sum |w_i|^2$ addition is called Regularization

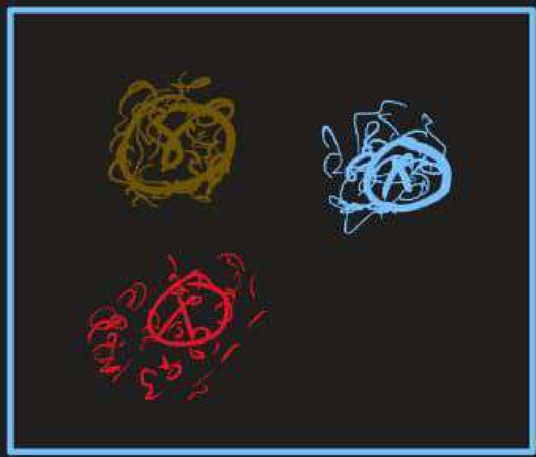
Kmeans++

14 December 2025 13:40

problem with normal Kmeans ⟹

its possible that the randomly initialized means are too close to each other

instead of



Kmeans++ ⟹ for the randomly chosen points

① choose 1 random pt. x_0 (distribute probability uniform)

② make probability s.t. $P(x_i) \propto \| x_i - x_0 \|^2$

so, new points will be further away !?

no closer possible

③ do it like that !

not used in the project