Machine Learning in Practice

Session 1 : Linear Regression

About me :

- Nikhil Bisen
- MTech - AI 2024 pass out from IISc Bangalore
- GATE 2022 - AIR 18(IN) , 233(EE)

Linear Regression : →

I) General Machine learning Set up : →

$n \rightarrow$ size of data

1) Data = (X_i, y_i)

① Data — $(x_i, y_i)_{i=1}$

x_i → features

y_i → target labels

② Model → function mapping from $x_i \to y_i$

$$x_i \to \boxed{\text{ML model}} \to \hat{y}_i$$

→ this function/model is characterized / parameterized by model weights (W)

$$x_i \to \boxed{W_1} \to \hat{y}_i$$

③ Loss function → tell us "how far we are from ground truth labels"

$$L(w) = \sum_{i=1}^{n} d(y_i, \hat{y}_i)$$

④ Optimization problem →

to find optimal weights

$$W^* = \underset{W}{\arg\min}\ L(w)$$

min $L(w)$

argmin $L(w)$ ↓ minima

⑤ ~~Gradient Descent~~ →

$$W_1 = W_0 - \alpha \nabla_{W_0} L(w)$$

$$W_2 = W_1 - \alpha \nabla_{W_1} L(w)$$

$\vdots$

W ↓ till convergence.

ω

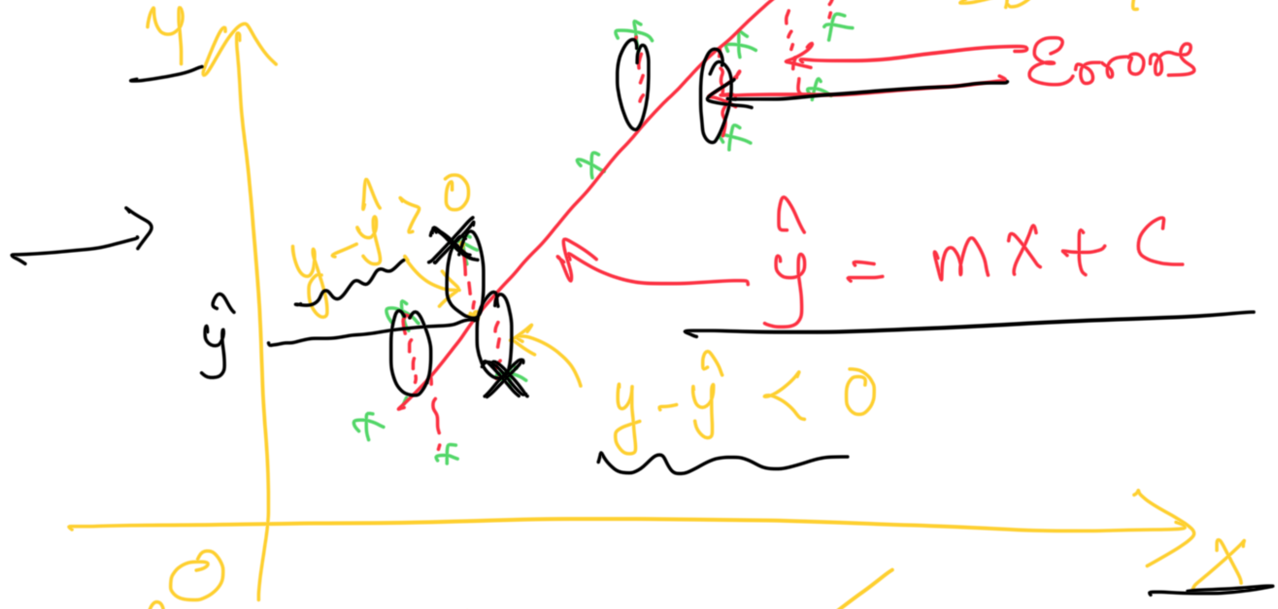
$x_i \in \mathbb{R}^2 = \begin{bmatrix} a \\ b \end{bmatrix}_{2\times 1}$

II) Simple linear regression :→

① Data ✓ $(x_i, y_i)_{i=1}^{n}$

$x_i \in \mathbb{R}$ (scalar) & $y_i \in \mathbb{R}$ (scalar)

② Model $\hat{y} = f(x) = mx + c$ ← a line in 2D space



③ Loss function :→ $\cancel{f(\cdot) = (\cdot)^2}$ ← differentiable

$f(\cdot) = |\cdot|$ ← Not differentiable

↳ MSE $= \frac{1}{N}\sum_{i=1}^{N}(\underbrace{y_i - \hat{y}_i}_{error_i})^2$

mean square error

④ Optimization problem

$$m^*, c^* = \arg\min \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]$$

m, c

But where is m & c !!

$$\hat{y}_i = m x_i + c$$

So,

$$m^*, c^* = \underset{m,c}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(y_i - (m x_i + c)\right)^2 \right]$$

Similarly (5) & ~~(6)~~ will be there.

III) Multiple linear Regression :→

(1) Data :→ $(x_i, y_i)_{i=1}^{n}$

But now $x_i \notin R$, $x_i \in R^d$

& $y_i \in R$

(2) model

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + \cdots\cdots + w_d x_{id}$$

I am ignoring the intercept

(3) loss function

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

(4) optim 2 problem

$$w_1^*, w_2^* \cdots w_n^* = \underset{w_1, w_2 \cdots w_n}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(y_i - (w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id})\right)^2 \right)$$

Lets write everything in compact form

1) Data :→ $(x_i, y_i)_{i=1}^{n}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$X = \begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_n — \end{bmatrix} \in \mathbb{R}^{n \times d}$$

$n \times d$

feature matrix

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

$x_i \in \mathbb{R}^{n \times 1}$

2) Model

$$\underline{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}_{d \times 1} \in \mathbb{R}^{d \times 1} \; ; \; \boxed{y_i = w^T \cdot x_i}$$

more compact :→ $\boxed{\hat{y} = X \cdot w}$

$y_1 = w^T x_1$

$y_2 = w^T x_2$

$\vdots$

$y_n = w^T x_n$

$$\begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_n — \end{bmatrix} \begin{bmatrix} | \\ w \\ | \end{bmatrix}$$

(3) $L = \|y - \hat{y}\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - w^T x_i)^2$

where, $\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$; $\hat{\underline{y}} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} \in \mathbb{R}^n$

$\|\cdot\|_2 \rightarrow$ L2 norm of vector

$\|v\|_2^2 = \frac{1}{N}(v^T \cdot v)$ for any vector v of length N.

$= \frac{1}{N}\sum_{i=1}^{N} v_i^2$

(4) optimization problem

$$w^* = \underset{w}{\operatorname{argmin}} \|y - X \cdot w\|_2^2$$

$(y - Xw) \in \mathbb{R}^n$

are this compatible ??

$X \in \mathbb{R}^{n \times d}$

$w \in \mathbb{R}^{d \times 1}$

$\|\cdot\|_2^2$

↳ L2-norm

$\|v\|_2^2 = \frac{1}{}v^T \cdot v$

N

$w \in \mathbb{R}^{n \times 1}$ ✓

$y \in \mathbb{R}$

⑤ Gradient descent :→

$$\frac{\partial L}{\partial w} = ?$$

$$\frac{\partial}{\partial w}\left(f(w) + g(w)\right) = \frac{\partial f(w)}{\partial w} + \frac{\partial g(w)}{\partial w}$$

$$\boxed{\frac{\partial L}{\partial w} = \frac{1}{N} \sum_{i=1}^{N} 2 \left(y_i - w^T x_i\right)\left(-x_i\right)}$$

$$= \frac{-2}{N} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)\left(x_i\right)$$

Now we can write iterative steps

Deriving Closed form Solution :→

$$\partial L / \partial w = 0$$

Some properties :

$A \in \mathbb{R}^{n\times n}$

$v \in \mathbb{R}^{n}$

1) $\frac{\partial}{\partial v}(v^T A v) = 2Av$

2) $\frac{\partial}{\partial v}(A \cdot v) = A^T$

3) $\frac{\partial}{\partial v}(v \cdot A) = A$

4) $(Av)^T = v^T \cdot A^T$

$L = \| y - Xw \|_2^2 = \frac{1}{N}(y - Xw)^T \cdot (y - Xw)$

dot product

$\|v\|_2^2 = \frac{1}{N} v^T \cdot v$

$L = (y^T - w^T X^T) \cdot (y - Xw)$

$= \frac{1}{N}\left(y^T y - y^T X w - w^T X^T y + w^T X^T X w\right)$

$\frac{\partial L}{\partial w} = \frac{1}{N}\left(0 - (y^T X)^T - X^T y + 2X^T X w\right)$

$= \frac{1}{N}\left(2X^T X w - 2X^T y\right)$

$\frac{\partial L}{\partial w} = 0 \Rightarrow \frac{2}{N} X^T y = \frac{2}{N} X^T X w^*$

$$\boxed{w^* = (\underbrace{X^TX})^{-1}(X^Ty)}$$

$\cancel{w^* = (X^TX)}$

IV) Ridge Regression :→

$$\underbrace{X^T}_{d\times n} \cdot \underbrace{X}_{n\times d} \Big\} d\times d$$

$$X^Ty = \underbrace{X^TX}\, w^*$$

$y = Ax$

$x = A^{-1}y$

$A^{-1}y = \underbrace{A^{-1}A}x$

$A^{-1}y = x$

$\cancel{1/A}$

$X \in \mathbb{R}^{n\times d}$

$X^TX \in \mathbb{R}^{d\times d}$

$$\frac{1}{2} \qquad \underline{A^{-1}}$$

$$w^* = \frac{X^Ty}{X^TX}$$

X^Ty ← ~~sea~~ vector

X^TX ← matrix

$$\begin{bmatrix}1\\2\end{bmatrix} \qquad \begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}$$

$$\frac{\begin{bmatrix}1\\2\end{bmatrix}}{\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}} \qquad \frac{\begin{bmatrix}1\\2\end{bmatrix}}{3} = \begin{bmatrix}1/3\\2/3\end{bmatrix}$$

$$\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix}^{-1} \cdot \begin{bmatrix}1\\2\end{bmatrix}$$

$$\begin{bmatrix} \quad \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$