08/08/24

MAP

MLE

(Maximum likelihood estimate).

n data points.

$D = \{x_1, x_2, \dots x_n\}$

likelihood of these data points, given a model $\theta = \{ \quad \}$

$$\log \hat{P}(D|\theta) = \hat{P}(\{x_1, x_2, \dots x_n\}|\theta) = \boxed{\prod_{k=1}^{n} \hat{P}(x_k|\theta)}$$

the MLE of θ is : $\theta_{MLE} = \arg\max_{\theta} \hat{P}(D|\theta)$

Coin flip

H | T

$$\boxed{\theta = P(H) = q = \frac{\# \text{ of } H}{\text{total Count}} = \frac{n_1}{n_1 + n_2}}$$

$D = \{x_1, x_2 \dots x_n\}$

$n_1 : H$

$n_2 : T$

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = q^{n_1}(1-q)^{n_2}$$

To obtain $\hat{q} = \arg\max_{q} q^{n_1}(1-q)^{n_2} = F$

$$\frac{\partial F}{\partial q} = n_1 q^{n_1 - 1}(1-q)^{n_2} - q^{n_1}(1-q)^{n_2 - 1} \times n_2 = 0$$

$$\Rightarrow q^{n_1 - 1}(1-q)^{n_2 - 1}(n_1(1-q) - q n_2) = 0$$

$$\Rightarrow q = \frac{n_1}{n_1 + n_2}$$

$$\log \hat{P}(D|\theta) = \sum_{i=1}^{N} \log(x_i|\theta) \qquad D = \{x_1 \dots x_n\}$$

MAP : Maximum a posteriori Estimate

$$\log(\theta_{MAP}) = \arg\max_{\theta} \underbrace{P(\theta|D)}_{\text{Posterior prob}} = \arg\max_{\theta} P(D|\theta)\, P(\theta)$$

likelihood, prior

$$= \log P(\theta) + \log P(D|\theta)$$

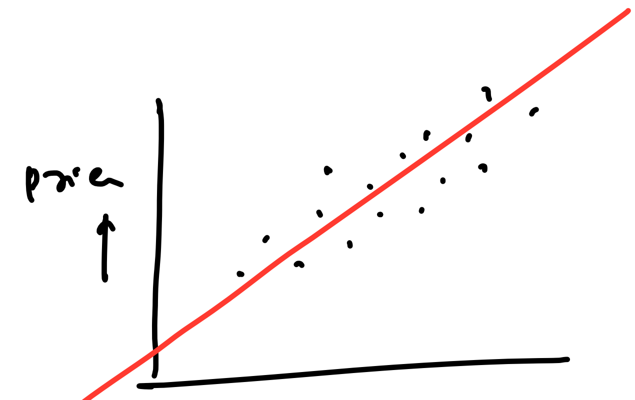
$$= \log P(\theta) + \sum_{i=1}^{n} P(x_i|\theta)$$

Linear Regression

Living area (Sqft)	price ($)
2104	400
1600	330
3000	232
→ 2500	?

$\| f(x) - y \|$

price ↑



$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$

$= \theta_0 + \sum_{i=1}^{n} \theta_i x_i$

x_0

$= \sum_{i=0}^{n} \theta_i x_i$

$x_0 = 1$

minimize the distance from all the points to the line L

$f(x_1 x_2 x_3) = x_1 + 3x_2 + 4x_3$

$(1, 2, 3)$

$f = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 =$

Loss: $\frac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2$

$\theta = \{\theta_0, \theta_1, \dots, \theta_n\}$

$d_i = (x_i, y_i)$

d_1
d_2
d_3
$\vdots$
d_m

$J(\theta) = \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$

m = # of training instance.

Superscript → instance

j = Subscript → = parameter / feature

$$\frac{\partial J}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_\theta(x^i) - y^i)^2$$

$$= 2 \times \frac{1}{2} (h_\theta(x^i) - y^i) \frac{\partial}{\partial \theta_j} (h_\theta(x) - y)$$

$$= (h_\theta(x^i) - y^i) \times \frac{\partial}{\partial \theta_j} \left(\sum_{k=1}^{n} \theta_k x_k - y \right)$$

$$= (h_\theta(x^i) - y^i) x_j$$

old

$\hat{\theta}_j \leftarrow \theta_j - \eta \frac{\partial J}{\partial \theta_j}$

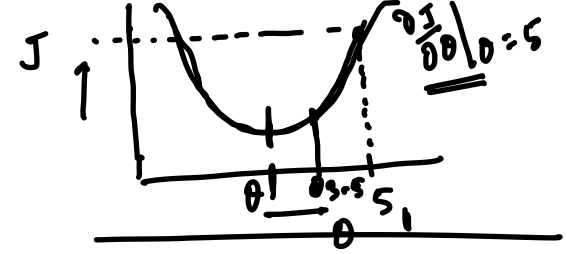
new

learning rate

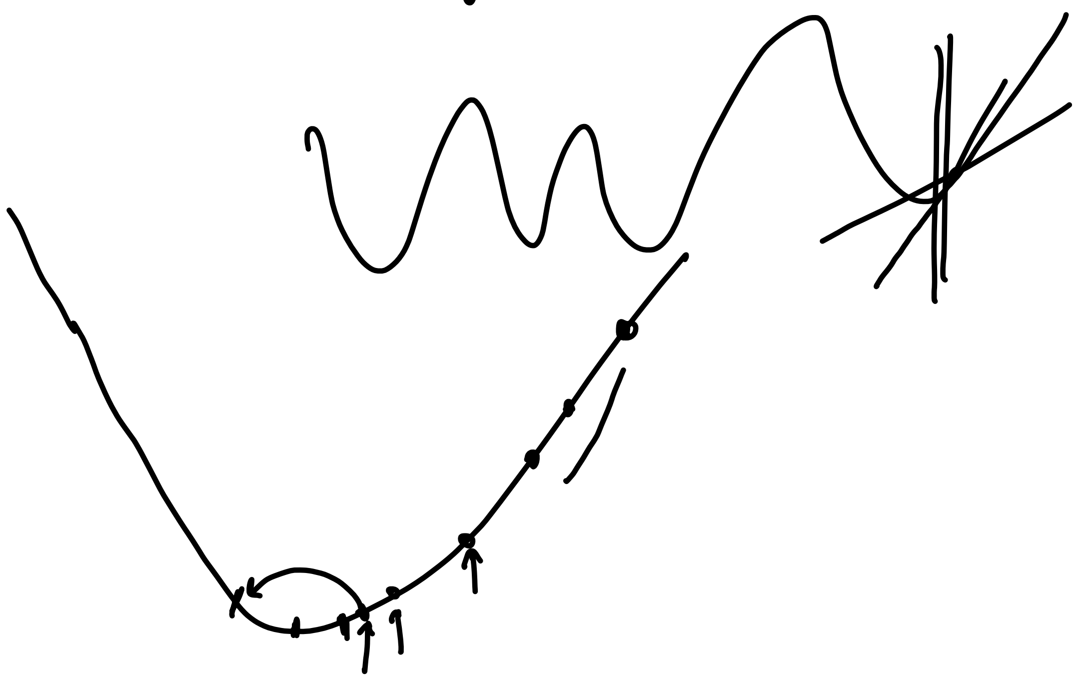
$$\frac{\partial J}{\partial \theta_p} = (h_\theta(x^i) - y^i) x_p$$

Batch GD | Stochastic GD | mini Batch GD

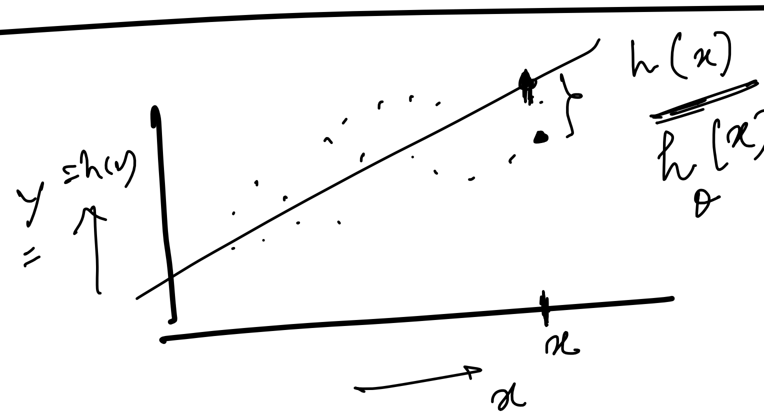
$$\hat{\theta}_j \leftarrow \theta_j - \eta \frac{\partial J}{\partial \theta_j}$$



$\theta = 5$



22/08/24



$$y = w_0 + w_1 x_1 + w_2 x_2 \cdots$$

θ

m = # of data points

n = # of features

$$\frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i \right)^2$$

$$x = \{ \; - - - - \; \}_n$$

The Normal Equation

$$f : \mathbb{R}^{m \times n} \to \mathbb{R}$$

$$f(A) = \frac{3}{2} A_{11} + 5 A_{12}^2 + A_{21} A_{22}$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} \end{bmatrix}$$

$$= \begin{bmatrix} 3/2 & 10 A_{12} \\ A_{22} & A_{21} \end{bmatrix}$$

$$\operatorname{tr}(A) = \operatorname{tr} A = \sum_{i=1}^{n} A_{ii}$$

1. $\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$
2. $\operatorname{tr}(A) = \operatorname{tr}(A^T)$
3. $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$
4. $\operatorname{tr}(aA) = a \operatorname{tr} A$

⑤ $\nabla_A \operatorname{tr}(AB) = B^T$ ⑥ $\nabla_{A^T} f(A) = (\nabla_A f(A))^T$

⑦ $\nabla_A \operatorname{tr} ABA^TC = CAB + C^TAB^T$

$$\vec{X} = \begin{bmatrix} — x^{(1)T} — \\ — x^{(2)T} — \\ — x^{(3)T} — \\ \vdots \\ — x^{(m)T} — \end{bmatrix}_{m \times n} \qquad \vec{Y} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix}_{m \times 1} \qquad \vec{\theta} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{n \times 1}$$

$$\underline{\underline{(\vec{X}\vec{\theta} - \vec{Y})}} = \begin{bmatrix} h_\theta(x^1) - y^1 \\ h_\theta(x^2) - y^2 \\ \vdots \\ h_\theta(x^m) - y^m \end{bmatrix} =$$

$$J(\theta) = \frac{1}{2}\underline{\underline{\sum (h_\theta(x^i) - y^i)^2}} = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

$$\nabla_\theta J(\theta) = \frac{1}{2}\nabla_\theta (X\theta - Y)^T(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T X^T - Y^T)(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y)$$

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta)\right]$$

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - \operatorname{tr}(\theta^T X^T Y) - \operatorname{tr}(Y^T X\theta)\right]$$

$A = Y^TX\theta$
$A^T = (Y^TX\theta)^T$
$= \theta^TX^TY$

$$= \frac{1}{2}\left[\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - 2\nabla_\theta \operatorname{tr}(Y^T X\theta)\right]$$

$$= \frac{1}{2}\left(\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta) - 2X^TY\right)$$

$\nabla_\theta \operatorname{tr}(Y^TX\theta)$
$= \nabla_\theta \operatorname{tr}(\theta Y^TX)$, $B = Y^TX$
$= (Y^TX)^T$

$\nabla_A \operatorname{tr}(AB) = B^T$

$\nabla_\theta \operatorname{tr}(\theta^T X^T X\theta)$, $A = \theta^T$, $B = X^TX$, $A^T = \theta$, $C = I$

$\cancel{\theta} = A^T$

$\nabla_\theta \operatorname{tr}(\theta\, \theta^T X^T X)$, $A = \theta$, $B = I$, $A^T = \theta^T$, $C = X^TX$

$= X^TX\theta + X^TX\theta$
$= 2X^TX\theta$

$\nabla_A \operatorname{tr} ABA^TC$
$= CAB + C^TAB^T$

$$\Rightarrow \frac{1}{2}\left(2X^TX\theta - 2X^TY\right) = 0$$

$$\boxed{X^TX\theta = X^TY} \longrightarrow \text{Normal Eq}$$

$$\theta = (X^TX)^{-1}X^TY \quad \text{if } X^TX \text{ is invertible}$$

$$y = w_0 + w_1 x_1 + \dots + w_n x_n$$

$$= w_0 + w_1 x_1^2 + w_2 x_2^2 + \dots + w_n x_n^2 \leftarrow$$

linear in the parameters.

$$Y = w_0 + w_1 \phi(x_1) + w_2 \phi(x_2) + \dots + w_n \phi(x_n)$$

ϕ = non-linear basis funct.

$\phi_j(x) = x^j$ – polynomial

$= \frac{(x-\mu)^2}{2\sigma^2}$ – Gaussian

$= \frac{1}{1+\exp(-s_i x_i)}$ – Sigmoid

$$J = \sum_{i=1}^{m}\left(y^i - \sum_{j=1}^{n}\theta_j \phi(x^i)\right)^2 = \sum\left(y^i - \theta^T \underbrace{\phi(x^i)}_{x^i}\right)^2$$

$$\frac{\partial J}{\partial \theta} = \sum\left(y^i - \theta^T x^i\right)^2 = 2\sum\left(y^i - \theta^T x^i\right)x^{(i)} = 0$$

$$\sum y^{(i)} x^{(i)T} = \theta^T \sum x^{(i)T} x^{(i)}$$

$$\Rightarrow X^TY = (X^TX)\theta \quad \text{Normal}$$

$$\Rightarrow \boxed{\theta = (X^TX)^{-1}(X^TY)}$$

$\cancel{X X^T}$ (X^TX) = non-invertible

$X \mid m \times n$

Regularizer

Ridge

L2

Ridge Regression

$$\hat{\theta} = \arg\min \sum_{i=1}^{m} (y_i - x_i \theta)^2 + \lambda \|\theta\|_2^2$$

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T Y$$

L1

Lasso Regression

$$\hat{\theta} = \sum_{i=1}^{m} (y_i - x_i \theta)^2 + \lambda \|\theta\|_1$$

Prove that : $\nabla_A \, tr(ABA^T C) = C^T A B^T + CAB$

→ we use two theorems as follows:

$tr(ABCD) = tr(DABC) = tr(CDAB) = tr(BCDA)$ — (I)

chain Rule of matrix differentiation:

$d(AB) = d(A)B + A\,d(B)$ — (II)

Therefore, $\nabla_A \, tr(ABA^T C) = \nabla_X \, tr(XBA^T C) + \nabla_X \, tr(ABX^T C)$

$\Rightarrow \nabla_X \, tr(XBA^T C) = (BA^T C)^T \quad (\because \nabla_A tr(AB) = B^T)$

$= C^T A B^T$ — (III)

$\nabla_X \, tr(ABX^T C) = \nabla_X tr(X^T CAB)$ (due to (I))

$\Rightarrow ((\nabla_X \, tr(X^T CAB)^T)^T$

as $\nabla_{A^T} f(A) = (\nabla_A f(A))^T \Rightarrow (\nabla_{X^T} \, tr(X^T CAB))^T$

as $\nabla_A tr(AB) = B^T \Rightarrow ((CAB)^T)^T$

$\Rightarrow CAB$ — (IV)

There from (III) and (IV)

$\nabla_A \, tr(ABA^T C) = C^T A B^T + CAB$