Deep learning lecture-2

$\nabla_x (Ax) = A^T$

$$A_{m\times n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & & \vdots \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$Ax = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \cdots a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots a_{mn}x_n \end{bmatrix}$$

To prove:

$\nabla_x (Ax) = A^T$

$\nabla_x (x^T A) = A$

$\nabla_x (x^T A x) = (A + A^T)x$

$\nabla_x (\|x\|^2) = 2x$

norm

$$\frac{dAx}{dx_1} = [a_{11} \ a_{21} \ \cdots \ a_{m1}]_{m\times 1}$$

$$\nabla_x (Ax) = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a & & \end{bmatrix}$$

$y = f(g(x))$

Univariate

$y'(x) = f'(g(x))\, g'(x)$

multivariate

$$y(x) = f(\bar{u}) \quad \text{vector}$$

$\bar{u} = g(\bar{x})$

$\bar{U} = [u_1 \ldots u_k] \quad x = [x_1 \ldots x_n]$

$$\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial u_1} \cdot \frac{\partial u_1}{\partial x_i} + \frac{\partial f}{\partial u_2} \frac{\partial u_2}{\partial x_i} + \ldots$$

$$\ldots \frac{\partial f}{\partial u_k} \cdot \frac{\partial u_k}{\partial x_i}$$

$\nabla_x y = A \nabla_u y$

$f(v^T x)$

where

$$A = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_n} \\ \frac{\partial u_2}{\partial x_1} & & \vdots \\ \vdots & & \\ \frac{\partial u_m}{\partial x_1} & & \frac{\partial u_m}{\partial x_n} \end{bmatrix}_{n \times m}$$

$$\nabla_u y = \begin{bmatrix} \frac{\partial y}{\partial u_1} \\ \vdots \\ \frac{\partial y}{\partial u_m} \end{bmatrix}_{m \times 1}$$

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Essential Job to do Phone No.

$\nabla_x y_1$

$\nabla_x y_2$ $\quad J = \begin{bmatrix} \partial y_1 \\ \\ \\ \end{bmatrix}$

$\vdots$

$\nabla_x y_N$

Pleatoric → inherent to the system

Epatemic → uncertainity due to sampling technique

$\frac{1}{\sqrt{n}}$

$\cancel{x}$ ~~ $\bar{\mu} = E \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$

$\Sigma = \sigma^2 = (x-\mu)^T (x-\mu)$

$\Sigma = \cancel{v^T \Sigma} \; V^T \Sigma V$

$\text{variance} = E\left[(f(x) - f(x))^2\right]$

$f(x) = b^T x$

$\text{Variance} = E\left[(f(x) - E(x))^2\right]$

$= E\left[(b^T x - E(b^T x))^2\right] = E\left[(b^T x - E(b^T x))(b^T x - E(b^T x)^T)\right]$

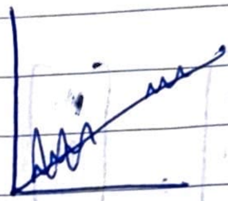
$= E\left[b^T (x - E(x))(x - E(x))^T b)\right]$

$= b^T ($

Linear Regression :-

Aleatoric → ~~yo~~ can model uncertainty that you know exists.

Linear Regression



Least count in physical measuring device

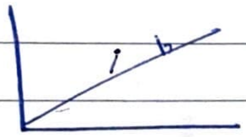
When to use:

① Linear relation

② Noise is well behaved

$$y = w^T x + b$$

$$\bar{y} = w^T \bar{x} + b$$



We use y distance not normal | perpendicular because it is easier to optimize

$$L = (\hat{y} - y)^2 \quad L = \sum_{i=1}^{n} (\hat{y}^i - y^i)^2$$

$$L(w) = \|Xw - \bar{y}\|^2$$

$$\nabla_w L(w) = 0$$

$$\nabla_x (x^T A x) = (A + A^T) x$$

$$\|x\|^2 = 2x$$

$$\nabla_x (A x) = A^T$$

$$(Xw - y)(2I) = 0$$

$$XX^T w - X^T y = 0$$

$$w^* = (XX^T)^{-1} X^T y$$

Essential	Job to do	Phone No.

Lab

Deep Learning Course #1

Sol^n 1.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

~~A = {~~ ~~f~~ $A = \{2, 4, 6\}$ $B = \{4, 6\}$

$(A \cap B) = \{4, 6\}$ $P(A \cap B) = \frac{2}{6} = \frac{1}{3}$

Sol^n 3. Let the element of A be a_{ij} and similarly the element of B be b_{ij}

therefore $A^T = a_{ji}$ and $B^T = b_{ji}$

LHS.

$(A^T + B^T) = a_{ji} + b_{ji}$

RHS.

$(A+B)^T = \cancel{A^T}\ (a_{ij} + b_{ij})^T$

$= a_{ji} + b_{ji}$

Hence, proved LHS = RHS

Essential | Job to do | Phone No.

$\underline{0,1,2}$

Sol'n.

[2 1 0 2 2 1 0 0

$$f(x) = 3x_1^2 + 5e^{x_2}$$

$$\nabla f(x_1, x_2) = \begin{bmatrix} \frac{\delta}{\delta x_1}(f) \\ \frac{\delta}{\delta x_2}(f) \end{bmatrix} = \begin{bmatrix} \cancel{6x_2} \; 6x_1 \\ 5e^{x_2} \end{bmatrix}$$

www.math.net/gradient

Sol 5.

$$u = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} \quad v = \begin{bmatrix} 2 \\ 4 \\ -1 \end{bmatrix} \quad u \cdot v = \begin{bmatrix} 6 \\ -4 \\ -2 \end{bmatrix}$$

Essential	Job to do	Phone No.

Deep learning

Linear Regression

$$L(w) = \frac{1}{N} \| \underbrace{X}_{n\times d}\underbrace{u}_{d\times 1} - \underbrace{y}_{n\times 1} \|^2$$

D+1

$$\nabla_w L(w) = 0$$

$$w^* = \underbrace{(X^T X)^{-1}}_{\text{not invertible}} (X^T y)$$

BSGD

Batch schotatic gradient descent

$$w, b = \langle w, b \rangle - \frac{\overset{\text{learning rate}}{\eta}}{\underset{\text{size of the batch}}{|B|}} \left(\partial_{w,b} (l(y, \hat{y})) \right)$$

$|B| = 1 \rightarrow$ stochastic (SGD)

SGD ($|B| = 1$) $|B|$ GD ($|B| = N$)

Batch size depends on your GPU memory

$$l(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{n} (w^T x + b - \hat{y})^2$$

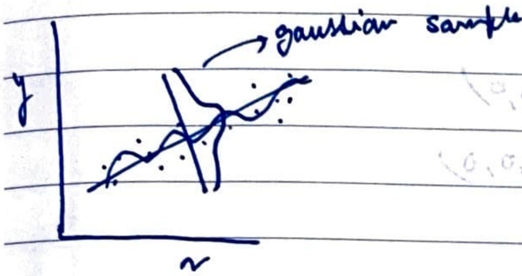
$$\frac{\partial l(y, \hat{y})}{\partial w} = \frac{1}{N} \sum_{i=1}^{n} 2 \left(w^T x^{(i)} + b - y^i \right) \left(2x^{(i)} \right)$$

$i \in B_t$

Log likelihood

$$P(Y|X) = \prod_{i=1}^{N} P(y^{(i)} | x^{(i)})$$

$$y = w^T x + b + \epsilon$$

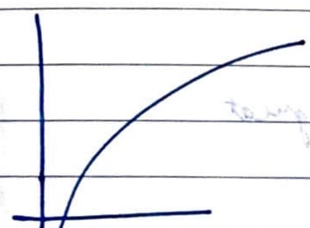


$$\epsilon \sim P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}$$

$$\mathcal{N}(\mu, \sigma^2) \qquad \epsilon = w^T x^{(i)} + b - y^{(i)}$$

$\mu \to 0$

$$P(y^i | x^i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(w^T x^{(i)} + b - y^{(i)} - 0)^2}{\sigma^2} \right\}$$



log function $$\log(P(y|x)) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\cancel{2}}{\cancel{2}} \sum_{i=1}^{N} \log\left(\frac{y^{(i)} - w^T x - b - 0}{\sigma^2}\right)^2$$

$$= \sum_{i=1}^{N} (y^{(i)} - w^T x^{(i)})^2$$

- add instead of multiplication

$\partial w, b, \sigma$ → Homoscedastic (uniform σ)
hetero " " (non-uniform σ)

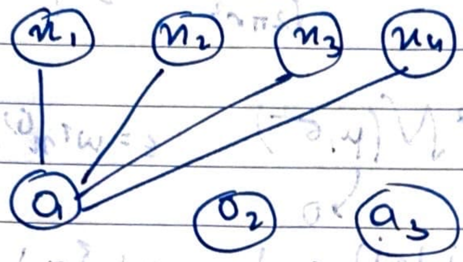
20 Saturday

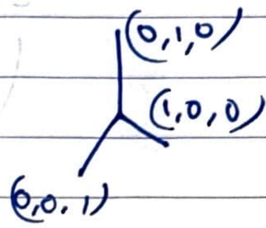
Linear classification

soft →

hard →

x_1	x_1
x_3	x_4





cat , dog , cow

$y^{(i)} =$ 1 2 3

ordinal regression

1, 2, 3 poor, average, great

$o_i = \sum w_j x_i + b_i$

21 Sunday

$$\frac{\exp\{o_i\}}{\sum_j \exp\{o_j\}}$$

$O = XW + b$ } n training, q categories, d input

$O = XW + b$

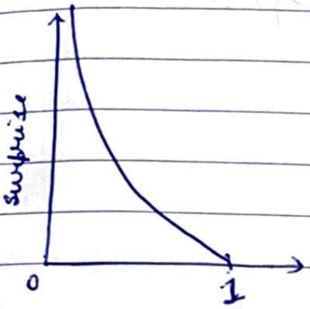
$$\begin{bmatrix} \ \end{bmatrix}_{n \times q} = \underset{n \times d}{\begin{bmatrix} X \end{bmatrix}} \underset{d \times q}{\begin{bmatrix} W \end{bmatrix}} + \begin{bmatrix} \ \end{bmatrix}_{n \times q}$$

Essential	Job to do	Phone No.

$P(\bar{Y}|\bar{X}) = \prod_{i=1}^{n} P(\bar{y}^i|\bar{x}^i)$

$-\log P(\bar{Y}|\bar{X}) = \sum_{i=1}^{n} -\log(P(\bar{y}^i|\bar{x}^i)) = \sum_{i=1}^{n} l(y^i, \hat{y}^i)$

$-\log P(\hat{y}^i)$ Expected surprise

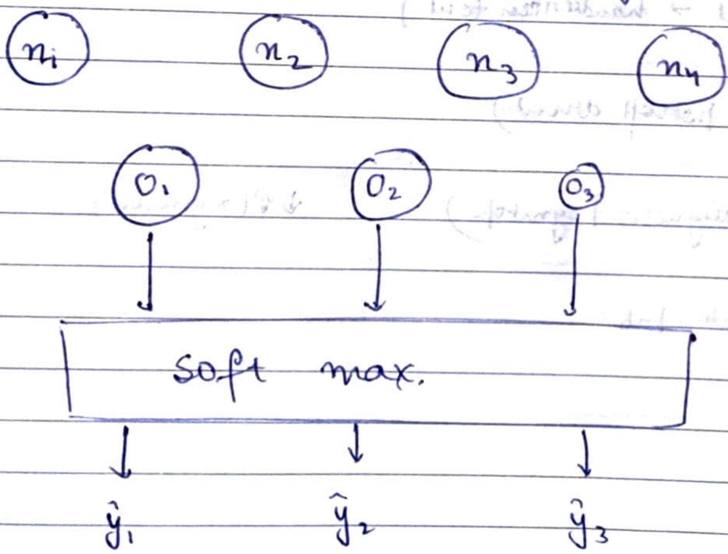


$\sum -\hat{y}\log P(y^i) \longrightarrow \frac{1}{N}\sum_{i=1}^{n} -y^i\log(\hat{y}^i)$

Entropy cross entropy

$y^i \to 1$ then $\log(\hat{y}^i) \to 0$ $\therefore \hat{y}^i \to 1$

$y^i \to 0$ then we need no change in log ~~to~~ $(\hat{y}^i)$



JANUARY

$$\begin{array}{r} 180 \\ +\ 50 \\ \hline 230 \end{array}$$

$\frac{230}{3} \times 2$

$$3 \overline{)460}$$ 153.3

3
16
15
10
9

$\frac{153}{230}$

2024

23 Tuesday

Cross Entropy loss

$\cancel{\frac{1}{N}} \cancel{\sum_{i=1}^{n}} \ l(\bar{y}$

PAC Probably Approx. correct

hypothesis

$P_{\mathcal{X}}\{|\epsilon(h) - \hat{\epsilon}(h)| \le \gamma\} > \delta$

→ training error

generalisation error

1. Empirical Error

$\epsilon_D(f) = \frac{1}{n}\sum_{i=1}^{n} 1\ (f(x^i) \neq y^i)$

net error

test set

2. Population Error

$D \sim p(x,y)$

Generalisation Error

1. Covariate Shift (printed text → handwritten text)
2. label shift
3. concept (beer in germany hotsoft drink)

Sympton given diagnosis P(diagnosis | syntom) ↓ P(digonosis)

and aumfamsonti. quarto. hub