ADVANCED COURSE IN MACHINE LEARNING : 1

① Frobenius norm for a $I \times J$ matrix : $\|X\|^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}^2 = \text{trace}(X^T X)$

a) and the derivative is $\frac{d}{dX} \|X\|^2 = 2X$. Prove it.

Let $f(X) = \|X\|^2$ for $f : \mathbb{R}^{i \times j} \longrightarrow \mathbb{R}$

Now we fix $(i,j) \Rightarrow \frac{df(X)}{dX_{ij}} = 2X_{ij} \Rightarrow \frac{df(X)}{dX} = 2X$

because $f(X) = X_{11}^2 + X_{12}^2 + \dots + X_{1J}^2 + \dots + X_{IJ}^2$ so if we fix (i,j) the derivative will be $2X_{ij}$.

b) Let $f(x) = (x+a)^T A^{-1}(Bx)$ where x is a column vector $x = x^{1 \times D}$

Compute the gradient $\nabla_x f(x)$.

- x is a vector column $D \times 1$ and a is the same.
- $(x+a)^T$ is a row vector $1 \times D$
- A, B are square matrices $D \times D$ $(\mathbb{R}^{D \times D})$
- Also $(x+a)^T A^{-1}$ is a row vector $1 \times D$ $(\mathbb{R}^{1 \times D})$
- Bx is a column vector $D \times 1$
- Finally $f(x) \in \mathbb{R}$

$f(x) = (x+a)^T A^{-1} Bx = x^T A^{-1} Bx + a^T A^{-1} Bx$

PROPERTIES : $\frac{d(x^T B x)}{dx} = x^T (B + B^T)$ and $\frac{d(a^T x)}{dx} = a^T$

so applying them : $\nabla_x f(x) = x^T (A^{-1}B + (A^{-1}B)^T) + a^T A^{-1} B$

Notice that $\nabla_x f(x) \in \mathbb{R}^{1 \times D}$ so it is a row vector.

② N data points $x_n \in \mathbb{R}^{D \times 1}$ with outputs y_n. Predict $\bar{y}_n = e^{x_n^T \theta + b}$ where $\theta \in \mathbb{R}^{D \times 1}$ and $b \in \mathbb{R}$.

We want to minimize the squared errors : $L = \sum_{n=1}^{N} w_n (y_n - \bar{y}_n)^2$ for $w_n > 0$.

a) Write the loss in matrix notation.

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} \dots x_{1D} \\ \vdots \ddots \vdots \\ x_{n1} \dots x_{nD} \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

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

$Y, \bar{Y} \in \mathbb{R}^{n \times 1}$

$\theta \in \mathbb{R}^{1 \times D}$

$$\bar{Y} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_n \end{pmatrix} = \begin{pmatrix} e^{x_1^T \theta + b} \\ \vdots \\ e^{x_n^T \theta + b} \end{pmatrix} = e^{X^T \theta + b}$$

Then the loss : $L = \| W^{\frac{1}{2}}(Y - \bar{Y}) \|^2 = \| W^{\frac{1}{2}}(Y - e^{x\theta+b}) \|^2$

b) Derivative wrt θ and b.

$$\frac{dL}{d\theta} = x(-e^{x\theta+b}) W 2 (W(Y - e^{x\theta+b}))$$

$$\frac{dL}{db} = (-e^{x\theta+b}) W 2 (W(Y - e^{x\theta+b}))$$

c) The gradient

$$\nabla_x L = \frac{dL}{dx} = \theta(-e^{x\theta+b}) W 2 (W(Y - e^{x\theta+b}))$$