Machine Learning 1 - Week 4

Nguyễn Thị Kiều Nhung - 11203041

October 05, 2022

Problem 1.

As before, we assume that the target variable t is given by a deterministic function $y(x, w)$ with additive Gaussian *noise* so that:

$$t = y(x, w) + \epsilon \tag{1}$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \text{ is an independent Gaussian noise}$$

you can see ϵ is a zero mean Gaussian random variable with precision (inverse variance) β or σ^2. Thus we can write:

$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1}) \tag{2}$$

In the case of Gaussian conditional distribution of the form (2), the conditional mean will be simply:

$$(t|x) = \int t\, p(t|x)\, dt = y(x, w) \tag{3}$$

Now consider a data set of input $X = \{x_1, x_2, ..., x_n\}$ with corresponding target value $\{t_1, t_2, ..., t_n\}$

$$X = \begin{bmatrix} — & x_{(1)}^T & — \\ — & x_{(2)}^T & — \\ & \vdots & \\ — & x_{(n)}^T & — \end{bmatrix};$$

$$\text{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}; \text{ w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}; \text{ y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix};$$

(4)

$$\implies t - y = \begin{bmatrix} t_1 - y_1 \\ t_2 - y_2 \\ \vdots \\ t_n - y_n \end{bmatrix}$$

$$\implies ||t_i - y_i||_2^2 = \sum_1^n (t_i - y_i)^2 = L$$

$$L = \|t - y\|_2^2 = \|t - wX\|_2^2 \tag{5}$$

(6)
$$\frac{\partial L}{\partial w} = 2X^T(t - Xw) = 0$$

$$\iff X^T t = X^T X w$$
$$\implies w = (X^T X)^{-1} X^T t$$

Besides, we have:

$$p(t|X, w, \beta) = \prod \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}) \tag{7}$$

We are not seeking to model the distribution of the input variable x. Thus x will always appear in the set of conditioning variables. So, we will drop explicitly x by taking the logarithm of the likelihood function:

$$\ln p(t|w, \beta) = \sum_{n=1}^{n} \ln \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}) = \frac{n}{2} \ln \beta - \frac{n}{2} \ln (2\pi - \beta E_D(w))$$

where the sum of squares error function is defined by:

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{n} \{t^n - w^T \phi(x^n)\}^2 \tag{8}$$

The gradient of the log likelihood function *(8)* takes the form:

$$\nabla \ln p(t|w, \beta) = \sum_{n=1}^{n} \{t_n - w^T \phi(x_n)\} \phi(x_n)^T \tag{9}$$

Set this gradient to zero:

$$0 = \sum_{n=1}^{n} t_n \phi(x_n)^T - w^T (\sum_{n=1}^{n} \phi(x_n) \phi(x_n)^T)$$

$$\begin{aligned} \implies w^T &= \frac{\sum_{n=1}^{n} t_n \phi(x_n)^T}{\sum_{n=1}^{n} \phi(x_n) \phi(x_n)^T)} \\ &= (\mathrm{X}^T X)^{-1} \cdot X^T t \end{aligned}$$

Problem 4. Prove $X^T X$ invertible when X full rank

Let X be the matrix $n \times k$. This matrix has a bunch of columns that are **linearly independent**

$$A = \begin{bmatrix} | & | & & | \\ \vec{a_1} & \vec{a_2} & \dots & \vec{a_n} \\ | & | & & | \end{bmatrix}; \tag{10}$$

All the columns of A $\vec{a_1}$, $\vec{a_2}$, ...,$\vec{_n}$ are linearLY independent. That mean only solution to

$$x_1\vec{a_1} + x_2\vec{a_2} + \ldots + x_n\vec{a_n} = 0$$

$$\implies \text{x}_i = 0$$
$$N(A) = \{\vec{0}\}$$

Let consider $A^T_{k\times n}A_{n\times k}$ is a $k \times k$ square matrix Where A is linearly independent columns
That mean the reduced row echelon form of the matrix $k \times k$ will have k pivot columns (there is only one $k \times k$ matrix with k pivot columns). We are mentioning the **"Identity matrix"** or invertible.

Let 's say have a vector $\vec{v}$, assume $\vec{v}_{k\times n} \in N(A^T A)$ means $(A^T A) \cdot \vec{v} = \vec{0}$

Multiply v^T at both side:

$$v^T_{1\times k}(A^T A)_{k\times k}\vec{v}_{k\times 1} = v^T \cdot \vec{0} = \vec{v} \cdot \vec{0} = 0 \tag{11}$$

On the LHS:

$$v^T_{1\times k}(A^T A)_{k\times k}\vec{v}_{k\times 1} = (A\vec{v})^T \cdot (A\vec{v}) = 0 \implies \|A\vec{v}\|^2 = 0 \leftrightarrow A\vec{v} = \vec{0} \tag{12}$$

$\implies \vec{v} \in N(A)(13)$

That is, $\vec{v}$ in both Null Space of $A^T A$ and Null Space of A implying:

$$N(A^T A) = N(A) = \{\vec{0}\}$$

Tell us, **only solution** to

$$(A^T A)\vec{x} = \vec{0}$$

is:

$$\vec{x} = \vec{0}$$

$\implies$ columns of $\text{A}^T A$ are Linearly Independent, reduced row echelon form of $A^T A$ is I_k (identity matrix)

$\implies \text{A}^T A$ is invertible.