Machine learning B
Assignment - 1

Mohit Gupta
MT22112

2. Linear Regression

(a) Pseudo-inverse → If A is a square matrix of order n, having rank r. The pseudo-inverse is a inverse of that matrix A, when the matrix A may not be invertible.

If matrix A is invertible, then its pseudo inverse is also equal to the simple matrix inverse. Matrix A can also be a $m \times n$ matrix.

The pseudo inverse is generally denoted by A^+

expression for pseudo-inverse for

i) Under-determined Solution of equations,

Under-determined systems are those systems having M equations, N variables &, $\boxed{M < N}$, & having multiple solutions

Let $Ax = b$

$$\boxed{x_{min} = A^T (AA^T)^{-1} b}$$

$$AA^+ = A^T A (A^T A)^{-1} = I$$

$$AA^+ = I$$

$$A^+ A \neq I$$

ii) Over-determined Solution of equations,

Over-determined system are those systems having M equations, N variables &, $\boxed{M > N}$,

for $Ax = b$.

$$\boxed{x_{min} = (A^T A)^{-1} A^T b}$$

$$A^+ A = (A^T A)^{-1} A^T A = I$$

$$A^+ A = I$$

$$AA^+ \neq I$$

b). Given system of linear equations,

$$x_1 + 3x_2 = 17$$
$$5x_1 + 7x_2 = 19$$
$$11x_1 + 13x_2 = 23$$

let's first convert this system into matrix form

$$\underset{3\times 2}{\begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} = \underset{3\times 1}{\begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}}$$

$$A \qquad X \qquad B$$

The equation is, $AX = B$

As we can see that, the number of variables (2) are less than the number of equations (3).

→ Now we have to make it equal

pre-multiply by A^T

$$A^T . A X = A^T B$$

$$\begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A^T \begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}$$

$$\underset{2\times 2}{\begin{bmatrix} 147 & 181 \\ 181 & 227 \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} = \underset{2\times 1}{\begin{bmatrix} 365 \\ 483 \end{bmatrix}}$$

→ Now we can observe that, we have same number of variables & equations.

$$X = A^{-1} . B$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{|A|} (\text{adj } A)\, B$$

$$= \frac{1}{(33369 - 32761)} \begin{bmatrix} 227 & -181 \\ -181 & 147 \end{bmatrix} \begin{bmatrix} 365 \\ 483 \end{bmatrix}$$

$$= \frac{1}{608} \begin{bmatrix} 82855 - 87423 \\ -66065 + 71001 \end{bmatrix}$$

Scanned with CamScanner

$$= \frac{1}{608}\begin{bmatrix} -4568 \\ 4936 \end{bmatrix} = \begin{bmatrix} -7.51 \\ 8.11 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -7.51 \\ 8.11 \end{bmatrix}$$

(c)

i) <u>Closed form expression</u>

Gradient descent = $\hat{y} = \theta^T . x$

where, $x \rightarrow$ input data

$y \rightarrow$ target variable

$\theta \rightarrow$ parameter

$\hat{y} \rightarrow$ prediction / hypothesis

Loss function = $\sum_{i=1}^{m} (y^i - \theta^T x^i)^2$

Normal equation is a closed-form solution for linear regression algorithm, which means we can obtain optimal parameter by just using a formula that includes few matrix multiplications & inversions

Normal equation,

$$\boxed{\theta = (X^T X)^{-1} X^T \vec{y}}$$

ii) We prefer iterative methods like gradient descent rather than using closed form solution to solve a linear regression problem, because the gradient descent is faster & less complex in computation.

The normal equation works on univariate cases, but when we have multiple variables the normal equation becomes much complex & requires more calculation.

3 Classification / Logistic Regression

(c) Gradient descent update rule

$$\theta_j := \theta_j - \alpha \frac{\partial L}{\partial \theta_j}$$

$\frac{\partial L}{\partial \theta_j} \rightarrow$ rate of change in loss function w.r.t θ_j

In logistic regression,

$$L(\theta) = -\sum_{i=1}^{n} y_i \cdot \log(h_\theta(x_i)) + (1-y_i)\log(1-h_\theta(x))$$

As, the derivative is linear, drop subscript i & compute for each training sample.

$$= -\frac{\partial}{\partial \theta_j}\left(y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))\right)$$

$$= -\left[y\frac{1}{h_\theta(x)} - (1-y)\frac{1}{1-h_\theta(x)}\right]\frac{\partial}{\partial \theta_j}(h_\theta(x)) \quad \text{——— (1)}$$

here, $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} = \tan h(x)$ (Sigmoid func^n)

$$1 - h_\theta(x) = 1 - \tanh(x)$$

$$\frac{\partial}{\partial \theta_j}(h_\theta(x)) = [1-\tanh^2(x)]\frac{\partial}{\partial \theta_j}(\theta^T x)$$

$$= [1-\tanh^2(x)]\, x^j$$

$$= [1-(h_\theta(x))^2]\, x^j \quad \text{——— (2)}$$

putting eq (2) in eq (1)

$$\frac{\partial L(\theta)}{\partial \theta_j} = -\left[y\frac{1}{h_\theta(x)} - (1-y)\frac{1}{1-h_\theta(x)}\right][1-(h_\theta(x))^2]\, x^j$$

$$= -\left[\frac{y(1-h_\theta(x)) - (1-y)h_\theta(x)}{h_\theta(x)\cancel{(1-h_\theta(x))}}\right](\cancel{1-h_\theta(x)})(1+h_\theta(x))\, x^j$$

$$= -\left[\frac{y - \cancel{y h_\theta(x)} - h_\theta(x) + \cancel{y h_\theta(x)}}{h_\theta(x)}\right](1+h_\theta(x))\, x^j$$

$$\boxed{\frac{\partial L}{\partial \theta_j} = -\sum_{i=1}^{n}\left[\frac{(y_i - h_\theta(x_i))(1+h_\theta(x_i))}{h_\theta(x_i)}\right]x_i^j}$$