COMP9417 – Machine Learning and Data Mining

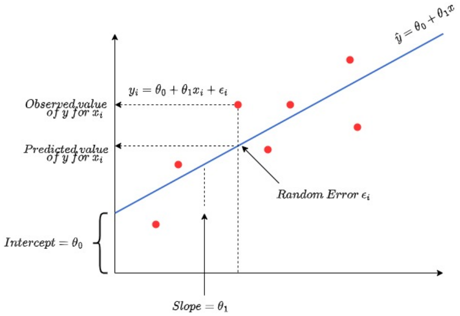
Tutorial Week 2: Linear Regression - 1

Arpit Kapoor
PhD Candidate
School of Mathematics and Statistics
arpit.kapoor@unsw.edu.au

February 22, 2023

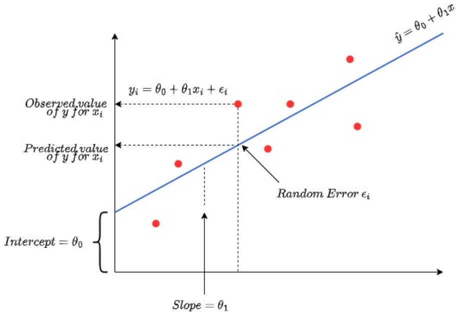
Linear Regression

- In regression problems, the target label y is real-valued.
- Assume a linear relationship between the output and feature(s)/input variable(s)
- This means the expected value of the output given input, $\mathbf{E}[y|x]$ is linear in input
- We can also say that the model output $\hat{y}$ is a linear combination of the input vector $\mathbf{x}$
- Training objective is to find values of the weights of this linear combination that minimize the loss function.



Linear Regression - Assumptions

- **Linearity:** The relationship between y and the mean of x is linear.
- **Homoscedasticity:** The variance of residual is the same for any value of x.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of x, y is normally distributed.



Linear Regression - Loss Function

- The sum of squared errorloss function for a univariate linear regression is written as,

$$J(\theta) = \sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2 \tag{1}$$

$$\text{where,} \quad \hat{y}_j = \theta_0 + \theta_1 x_j \tag{2}$$

- For multivariate/multiple linear regression,

$$\hat{y}_j = \theta_0 + \theta_1 x_{j1} + \theta_2 x_{j2} + \ldots + \theta_{p-1} x_{j(p-1)} \tag{3}$$

$$\hat{y}_j = \sum_{i=0}^{p-1} \theta_i x_{ji} \qquad \qquad where, x_{j0} = 1 \tag{4}$$

Linear Regression - Training

Find the values of θ that minimizes the cost function $J(\theta)$:

$$\hat{\theta} = \arg\min_{\theta} J(\theta) \tag{5}$$

We can approach this with the following ways:

- Gradient Descent
  - Iteratively update the parameters until the minima are achieved
  - Achieve this by taking small steps in the direction opposite to the gradient of the loss function w.r.t. the parameters
- Least-Square estimates
  - SSE loss function is a convex function
  - Analytically compute the global minima as the point where the gradient is 0
  - In the case of multiple linear regression, this can be generalized to the normal equation

Gradient Descent

Gradient Descent is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function. This method is commonly used in machine learning (ML) and deep learning(DL) to minimise a cost/loss function (e.g. in linear regression)

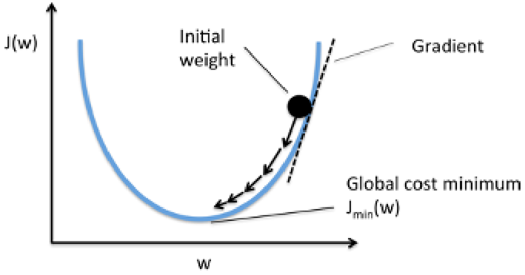


Figure: Gradient Descent optimization

Gradient Descent

- Start with initial weights (randomly assigned), and take small iterative steps towards local optima:

$$\theta_i^{t+1} := \theta_i^t - \alpha \nabla_{\theta_i}(J(\theta)) \tag{6}$$

- Here, α is called the *learning rate* used to control the size of each step in GD
- This is repeated for a fixed number of *epochs* or until a threshold value of $J(\theta)$ is reached
- In Batch gradient descent, we calculate the gradients for each batch containing m data samples,

$$\theta_i^{t+1} := \theta_i^t - \alpha \frac{1}{m} \sum_{j=1}^{m} \nabla_{\theta_i} \left(y_j - \hat{y}_j \right)^2 \tag{7}$$

Tutorial Questions

Least Square Estimate - Univariate Regression

Question 1 (a). Given the linear regression line,

$$y = \theta_0 + \theta_1 x \tag{8}$$

The mean squared error loss function for *n* training samples is written as,

$$\begin{aligned} J(\theta_0, \theta_1) &= \frac{1}{n}\sum_{j=1}^{n}(y_j - (\theta_0 + \theta_1 x_j))^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}(y_j - \theta_0 - \theta_1 x_j)^2 \end{aligned}$$

Least Square Estimate - Univariate Regression

Now, at the minimum value of J, the partial derivative of J w. r. t. θ_0 and θ_1 is 0,

$$\begin{aligned}
\frac{\partial J}{\partial \theta_0} &= \frac{1}{n}\sum_{j=1}^{n}\frac{\partial (y_j - \theta_0 - \theta_1 x_j)^2}{\partial \theta_0} \\
&= \frac{1}{n}\sum_{j=1}^{n} -2(y_j - \theta_0 - \theta_1 x_j) \\
&= -2\Big[\frac{1}{n}\sum_{j=1}^{n} y_j - \frac{\theta_0}{n}\sum_{j=1}^{n} 1 - \frac{\theta_1}{n}\sum_{j=1}^{n} x_j\Big] \\
&= -2(\overline{y} - \theta_0 - \theta_1 \overline{x}) \qquad (9)
\end{aligned}$$

On setting the partial derivative from the last step to 0 we get,

$$\theta_0 = \overline{y} - \theta_1 \overline{x} \qquad (10)$$

Least Square Estimate - Univariate Regression

Similarly, we can compute the partial derivative of the loss with respect to θ_1

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \frac{1}{n}\sum_{j=1}^{n}\frac{\partial (y_j - \theta_0 - \theta_1 x_j)^2}{\partial \theta_1} \\
&= \frac{1}{n}\sum_{j=1}^{n} -2(y_j - \theta_0 - \theta_1 x_j)x_j \\
&= -2\Big[\frac{1}{n}\sum_{j=1}^{n} x_j y_j - \frac{\theta_0}{n}\sum_{j=1}^{n} x_j - \frac{\theta_1}{n}\sum_{j=1}^{n} x_j^2\Big] \\
&= -2(\overline{xy} - \theta_0 \overline{x} - \theta_1 \overline{x^2}) \qquad (11)
\end{aligned}$$

On setting to 0, we get

$$\theta_1 = \frac{\overline{xy} - \theta_0 \overline{x}}{\overline{x^2}} \qquad (12)$$

Least Square Estimate - Univariate Regression

On substituting θ_0 from Eq 9, we get

$$\theta_1 = \frac{\overline{xy} - (\overline{y} - \theta_1 \overline{x})\overline{x}}{\overline{x^2}}$$

Finally, we have

$$\boxed{\theta_1 = \frac{\overline{xy} - \bar{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}} \tag{13}$$

And,

$$\boxed{\theta_0 = \overline{y} - \theta_1 \overline{x}} \tag{14}$$

Least Square Estimate - Univariate Regression

Question 1 (b). Show that the centroid point $(\overline{x}, \overline{y})$ is always on the least square regression line.
Our least squares regression line is:

$$\begin{aligned} \hat{y}(x) &= \hat{\theta}_0 + \hat{\theta}_1 x \\ &= \left(\overline{y} - \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \overline{x}^2}\overline{x}\right) + \left(\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \overline{x}^2}\right)x \end{aligned} \tag{15}$$

where the $\hat{y}$ is the predicted value and with estimated parameter values $(\hat{\theta}_0, \hat{\theta}_1)$.
Therefore, $\hat{y}(x)$ is our estimate of y at evaluation point x.
On substituting $x = \overline{x}$, we get

$$\hat{y}(x) = \overline{y}$$

So, our least squares regression line passes through $(\overline{x}, \overline{y})$

Least Square Estimate - Univariate Regression

Question 1 (c). Find least-squares estimate for L2-regularised linear regression with following loss function:

$$J(\theta_0, \theta_1) = \frac{1}{n}\sum_{j=1}^{n}(y_j - (\theta_0 + \theta_1 x_j))^2 + \lambda\theta_1^2$$

Solution. On repeating steps from Q.1(a) with the updated loss function we get:

$$\begin{aligned}\theta_0 &= \overline{y} - \theta_1\overline{x} \\ \theta_1 &= \frac{\overline{xy} - \bar{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2 + \lambda}\end{aligned}$$

Least Square Estimate - Multivariate Regression

Question 2. Given Design matrix X of shape $(n \times p)$ with $p-1$ input features for n samples and target vector $\mathbf{y}$ of shape $(n \times 1)$,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \quad \text{and,} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{16}$$

The regression parameters are written as,

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{p-1} \end{bmatrix} \tag{17}$$

Least Square Estimate - Multivariate Regression

The loss function is written as,

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 && (18)\\
&= (\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta}) && (19)
\end{aligned}
$$

$$
\begin{aligned}
\text{Let,} \qquad u &= (\mathbf{y} - X\boldsymbol{\theta}) && (20)\\
\text{So,} \qquad J(\boldsymbol{\theta}) &= g(u) = u^T u && (21)
\end{aligned}
$$

Using chain rule, the partial differential of $J(\boldsymbol{\theta})$ w. r. t. $\boldsymbol{\theta}$ can be written as follows,

$$
\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial g(u)}{\partial \boldsymbol{\theta}} = \frac{\partial g(u)}{\partial u} . \frac{\partial u}{\partial \boldsymbol{\theta}} \tag{22}
$$

Least Square Estimate - Multivariate Regression

Computing the partial derivatives,

$$\frac{\partial g(u)}{\partial u} = 2u^T; \qquad \text{and} \qquad \frac{\partial u}{\partial \boldsymbol{\theta}} = -X$$

On multiplying, we get

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= -2u^T X \\ &= -2(\mathbf{y} - \boldsymbol{\theta})^T X \end{aligned} \tag{23}$$

Setting the partial derivative to 0,

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= 0 \\ -2(\mathbf{y} - \boldsymbol{\theta})^T X &= 0 \\ -2X^T(\mathbf{y} - X\boldsymbol{\theta}) &= 0, \qquad [\text{using } (A.B)^T = B^T.A^T] \end{aligned} \tag{24}$$

Least Square Estimate - Multivariate Regression

$$
\begin{aligned}
-X^T\mathbf{y} + X^T X\boldsymbol{\theta} &= 0 \qquad (25)\\
X^T X\boldsymbol{\theta} &= X^T\mathbf{y} \qquad (26)
\end{aligned}
$$

From the above solution, the critical point value of $\boldsymbol{\theta}$ is given as,

$$\boxed{\boldsymbol{\theta} = (X^T X)^{-1} X^T\mathbf{y}} \qquad (27)$$

Note: It is assumed that the matrix $X^T X$ is invertible

Question 2 (b). To prove that the critical point obtainedis the global optima, we will compute the Hessian of $J(\boldsymbol{\theta})$ and show that it is positive semi-definite, Hessian of loss $J(\boldsymbol{\theta})$ is written as,

$$
\begin{aligned}
H = \nabla^2_\theta J(\boldsymbol{\theta}) &= \nabla_\theta(\nabla_\theta J(\boldsymbol{\theta}))\\
&= \nabla_\theta(-2X^T y + 2X^T X\boldsymbol{\theta})\\
&= 2X^T X \qquad (28)
\end{aligned}
$$

Least Square Estimate - Multivariate Regression

Matrix H is positive semi-definite if for any vector $\mathbf{u} \in \mathbb{R}^p$,

$$\mathbf{u}^T H \mathbf{u} \geq 0$$

We start by computing the left-hand side,

$$\begin{aligned} \mathbf{u}^T H \mathbf{u} &= \mathbf{u}^T (2X^T X)\mathbf{u} \\ &= 2(\mathbf{u}^T X^T)(X\mathbf{u}) \\ &= \|X\mathbf{u}\|_2^2 \geq 0 \end{aligned} \tag{29}$$

We know that the squared euclidean norm is always positive. Therefore, J is a convex function and the critical point is the global optima.

Least Square Estimate - Multivariate Regression

Question 2(c). The single input sample of the univariate linear regression problem is written as,

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_{i1} \end{bmatrix} \tag{30}$$

And the target vector is written as,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \tag{31}$$

The parameter vector **w** is written as,

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in \mathbb{R}^2 \tag{32}$$

Least Square Estimate - Multivariate Regression

Each row in the design matrix, X is the transpose of the input vector $\mathbf{x}_i$,

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \tag{33}$$

and,

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} = \begin{bmatrix} n & n\overline{x} \\ n\overline{x} & n\overline{x^2} \end{bmatrix} \in \mathbb{R}^{2\times 2} \tag{34}$$

Least Square Estimate - Multivariate Regression

For a 2×2 matrix of the form,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \text{ then, } A^{-1} = \frac{1}{(ad - bc)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{35}$$

So,

$$(X^T X)^{-1} = \frac{1}{(n^2 \overline{x^2} - n^2 \overline{x}^2)} \begin{bmatrix} n\overline{x^2} & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \tag{36}$$

$$= \frac{1}{n(\overline{x^2} - \overline{x}^2)} \begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \tag{37}$$

Finally,

$$X^T \mathbf{y} = \begin{bmatrix} n\overline{y} \\ n\overline{xy} \end{bmatrix} \tag{38}$$