CS725: Foundations of Machine Learning

Linear Regression (Contd)

August 24, 2020

Recall: Closed Form Solution for Linear Regression (1D)

Consider the simplest case, i.e. 1D data, $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$.

$\hat{y}_i = w_0 + w_1 x_i$

Objective: Find $w_0^*, w_1^* = \arg\min_{w_0, w_1} E(w_0, w_1, \mathcal{D}) = \arg\min_{w_0, w_1} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$

Solving $(w_0^*, w_1^*) = \arg\min_{w_0, w_1} \sum_{i=1}^{m} (y_i - w_0 - w_1 x_i)^2$ gives us

$$w_0^* = \bar{y} - w_1 \bar{x}$$

and

$$w_1^* = \frac{\sum_i \frac{x_i y_i}{n} - \bar{x}\bar{y}}{\sum_i \frac{x_i^2}{n} - \bar{x}^2}$$

Gradient

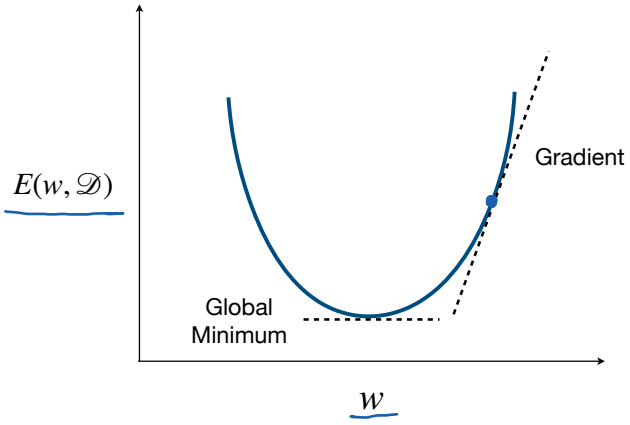


Figure 1: Illustration of gradients/derivatives in 1D

Closed Form Solution for d-dimensional data

More generally, each $\mathbf{x}_i$ is a d-dimensional vector, $\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$. $\in \mathbb{R}$, $d \times 1$

Closed Form Solution for d-dimensional data

More generally, each $\mathbf{x}_i$ is a d-dimensional vector, $\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$.

For d-dimensional data, the matrix $\mathbf{X}$ can be defined as:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots\ldots & x_{1d} \\ \vdots & & & \\ x_{n1} & x_{n2} & \ldots\ldots & x_{nd} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \quad (1)$$

$n \times d$

and the output $\mathbf{y}$ is a column vector, $\mathrm{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Closed Form Solution for d-dimensional data

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}_{d \times 1} \qquad x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}_{d \times 1}$$

$$\underset{w}{\operatorname{argmin}} \; \|y - Xw\|_2^2$$

Closed Form Solution for d-dimensional data

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Solve for $\mathbf{w}$ by setting $\nabla_{\mathbf{w}} E = \nabla_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 = 0$ and solving for $\mathbf{w}$.

$$\begin{bmatrix} \frac{\partial E}{\partial w_1} \\ \vdots \\ \frac{\partial E}{\partial w_d} \end{bmatrix}$$

Closed Form Solution for d-dimensional data

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Solve for $\mathbf{w}$ by setting $\nabla_{\mathbf{w}} E = \nabla_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 = 0$ and solving for $\mathbf{w}$.

$$\nabla_w E = -2 \sum_i (y_i - w^T x_i) x_i = 0$$

$$\Rightarrow -2 \sum_i y_i x_i + 2 \sum_i (w^T x_i) x_i = 0$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}_{n \times d} \quad \nabla_{\mathbf{w}} E = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \mathbf{w} = 0 \Rightarrow \boxed{\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}_{d \times 1}$$

Closed Form Solution for d-dimensional data

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w}, \mathcal{D}) = \arg\min_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Solve for $\mathbf{w}$ by setting $\nabla_{\mathbf{w}} E = \nabla_{\mathbf{w}} \sum_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 = 0$ and solving for $\mathbf{w}$.

$$\nabla_{\mathbf{w}} E = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \Rightarrow \boxed{\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

Another way to derive this solution: Using geometrical intuition!

- The minimum value attainable for the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, where

$$\Phi = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nd} \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

$\operatorname{argmin}_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$

- The minimum value attainable for the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have $\mathbf{X}\mathbf{w}^* = \mathbf{y}$, where

$$\Phi = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nd} \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

- It has a solution if $\mathbf{y}$ is in the column space of $\mathbf{X}$.

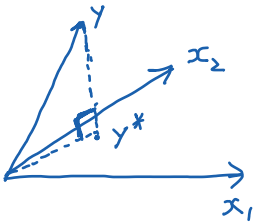
- If $\mathbf{y}$ is NOT in the column space of $\mathbf{X}$ (that is, if zero is NOT attainable at $\mathbf{w}^*$), what can be done?

Geometric Interpretation of Least Square Solution

- Let $\mathbf{y}^*$ be a solution in the column space of $\mathbf{X}$
- The least squares solution is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized

Geometric Interpretation of Least Square Solution

- Let $\mathbf{y}^*$ be a solution in the column space of $\mathbf{X}$
- The least squares solution $\mathbf{w}^*$ is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized
- Therefore, the line joining $\mathbf{y}^*$ to $\mathbf{y}$ should be orthogonal to the column space



$$(\mathbf{y} - \mathbf{y}^*)^T \mathbf{X} = 0$$
$$(\mathbf{y}^*)^T \mathbf{X} = (\mathbf{y})^T \mathbf{X}$$

$$(\mathbf{X}\mathbf{w}^*)^T\mathbf{X} = \mathbf{y}^T\mathbf{X} \qquad \text{(Substituting } \mathbf{X}\mathbf{w}^* = \mathbf{y}^*\text{)}$$

$(AB)^T = B^T A^T$

$$\mathbf{w}^{*T}\mathbf{X}^T\mathbf{X} = \mathbf{y}^T\mathbf{X}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w}^* = \mathbf{X}^T\mathbf{y}$$

$$\boxed{\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}$$

$y^* = Xw^* = X(X^TX)^{-1}X^Ty$

$y^* - y = \left(X(X^TX)^{-1}X^T - I\right)y$

$X^T(y^* - y) =$

$\left(X^TX(X^TX)^{-1}X^T - X^T\right)y$

$= 0$!

Inverse in the Closed Form Solution

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

Here, $\mathbf{X}^T\mathbf{X}$ is invertible if $\mathbf{X}$ has full column rank.

Proof :

Assume that $\mathbf{X}$ has full column rank. There exists $\mathbf{w}$ such that $\mathbf{X}^T\mathbf{X}\mathbf{w} = 0$.

$$\begin{aligned} &\Rightarrow \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \\ &\Rightarrow (\mathbf{X}\mathbf{w})^T\mathbf{X}\mathbf{w} = 0 \\ &\Rightarrow \mathbf{X}\mathbf{w} = 0 \\ \Rightarrow \mathbf{w} = 0 &\text{ (since } \mathbf{X} \text{ is full column rank)} \end{aligned}$$

Hence, $\mathbf{X}^T\mathbf{X}$ is invertible if $\mathbf{X}$ is full column rank

Regression Model with Basis Functions

For 1D data,

$$\hat{y}_i = w_0 + w_1 x_i + (w_2 x_i^2 + w_3 x_i^3 + \ldots + w_m x_i^m) = \sum_{j=0}^{m} w_j \phi_j(x_i)$$

where $\phi_j(x)$ is called a **basis function**.

$\phi_0(x_i) = 1$

$\phi_1(x_i) = x_i$

$\phi_2(x_i) = x_i^2$

Regression Model with Basis Functions

For 1D data,

$$\hat{y}_i = w_0 + w_1 x_i + (w_2 x_i^2 + w_3 x_i^3 + \ldots + w_m x_i^m) = \sum_{j=0}^{m} w_j \phi_j(x_i)$$

where $\phi_j(x)$ is called a **basis function**.

Common basis functions (for 1D data):

- Polynomial basis: $\phi_j(x) = x^j$
- Radial basis function: $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2\sigma^2}}$

mean of the j^{th} basis function

More generally, a regression model can be written using basis functions as:

$$\hat{y}_i = w_0 \phi_0(\mathbf{x}_i) + w_1 \phi_1(\mathbf{x}_i) + \ldots + w_m \phi_m(\mathbf{x}_i) = \sum_{j=0}^{m} w_j \phi_j(\mathbf{x}_i)$$

$m > d$ $(\mathbf{x}_i \in \mathbb{R}^d)$

Regression Model with Basis Functions

For 1D data,

$$\hat{y}_i = w_0 + w_1 x_i + (w_2 x_i^2 + w_3 x_i^3 + \ldots + w_m x_i^m) = \sum_{j=0}^{m} w_j \phi_j(x_i)$$

where $\phi_j(x)$ is called a **basis function**.

Common basis functions (for 1D data):

- Polynomial basis: $\phi_j(x) = x^j$
- Radial basis function: $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2\sigma^2}}$

More generally, a regression model can be written using basis functions as:

$$\hat{y}_i = w_0 \phi_0(\mathbf{x}_i) + w_1 \phi_1(\mathbf{x}_i) + \ldots + w_m \phi_m(\mathbf{x}_i) = \sum_{j=0}^{m} w_j \phi_j(\mathbf{x}_i)$$

Regression Model with Basis Functions

Objective: Find $\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}_i))^2$ where $\mathbf{\Phi}(\mathbf{x}_i) = \begin{bmatrix} \phi_0(\mathbf{x}_i) \\ \vdots \\ \phi_m(\mathbf{x}_i) \end{bmatrix}$

Closed form solution: $\boxed{\mathbf{w}^* = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{y}}$

where $\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}(\mathbf{x}_1)^T \\ \vdots \\ \mathbf{\Phi}(\mathbf{x}_n)^T \end{bmatrix}$