DDA3020 Week 3 Tutorial

Linear Regression - Part 1

Kangran ZHAO

School of Data Science,
The Chinese University of Hong Kong, Shenzhen

February 9, 2025

INTRODUCTION

- Given data $(\mathbf{x}, y)$ sampled from random variable $\mathcal{X}$ and other random variable(s) $\mathcal{Y}$, we need to find out a linear relationship between them, e.g., the linear relationship between the price of a house and the size of a house.
- One variable linear relationship:

$$y = b + wx$$

- Many variable linear relationship:

$$\begin{aligned} y &= w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_k x_k \\ &= w_0 + \sum_{i=1}^{k} w_i x_i, (\text{where } x_0 = 1) \\ &= \mathbf{x}^T \mathbf{w} \end{aligned}$$

- The goal of linear regression is to find out the optimal value of **a** so to minimize the gap between your predicted value and actual value. That is, given data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, the value you predicted using $\hat{y}_i = \mathbf{x}^T \mathbf{w} + b$ should be as close to actual value of y_i. In linear regression, the goal is to minimize the average prediction error

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{x}_i^T \mathbf{w} - y_i\right)^2 \\ =&\frac{1}{m} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) \end{aligned}$$

Closed-Form Solution

- Solution (from lecture):

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} \frac{1}{m} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) &= 0 \\
\frac{\partial}{\partial \mathbf{w}} \left(\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{y}^T \mathbf{y} \right) &= 0 \\
2\mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{X}^T \mathbf{y} &= 0 \\
2\mathbf{X}^T \mathbf{X}\mathbf{w} &= 2\mathbf{X}^T \mathbf{y} \\
\hat{\mathbf{w}} &= \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}
\end{aligned}
$$

- Supplementary about vector/matrix derivatives:

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^T \mathbf{x} &= \begin{bmatrix} \frac{\partial}{\partial x_1} \mathbf{a}^T \mathbf{x} & \dots & \frac{\partial}{\partial x_m} \mathbf{a}^T \mathbf{x} \end{bmatrix}^T \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_{i=1}^k a_i x_i & \dots & \frac{\partial}{\partial x_m} \sum_{i=1}^k a_i x_i \end{bmatrix}^T \\
&= \begin{bmatrix} a_1, \dots, a_m \end{bmatrix}^T \\
&= \mathbf{a}
\end{aligned}
$$

- Small Exercise

$$
\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = \begin{bmatrix} \frac{\partial}{\partial x_1} \mathbf{x}^T \mathbf{A} \mathbf{x} & \dots & \frac{\partial}{\partial x_m} \mathbf{x}^T \mathbf{A} \mathbf{x} \end{bmatrix}^T = \dots = \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{x}
$$

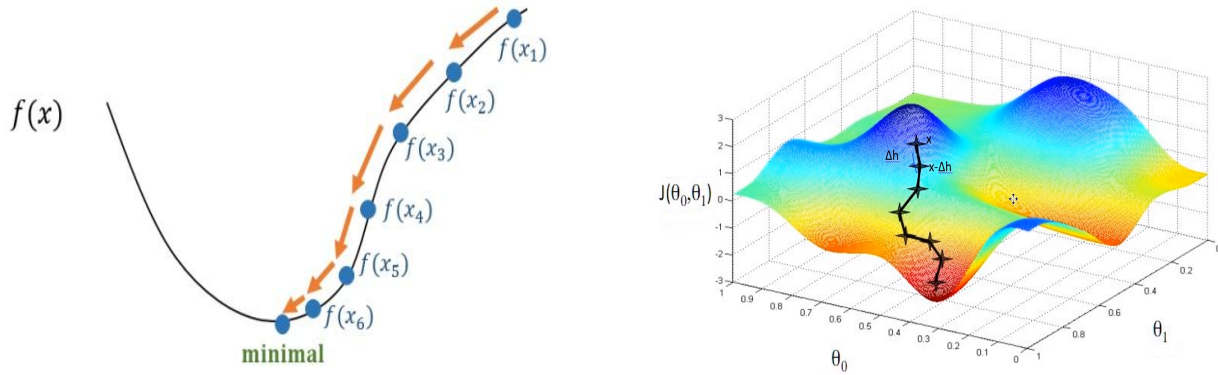
Extension of Matrix/Vector/Scaler Derivatives

- Usually we have nominator layout and denominator layout, which means to mainly use nominator or denominator matrix/vector/constant shape as output. Take denominator layout as example:

	$y \in \mathbb{R}$	$\mathbf{y} \in \mathbb{R}^p$	$\mathbf{Y} \in \mathbb{R}^{p \times q}$
$x \in \mathbb{R}$	$\mathbb{R}$	$\begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial x} & \cdots & \frac{\partial \mathbf{y}_p}{\partial x} \end{bmatrix} \in \mathbb{R}^{1 \times p}$	$\begin{bmatrix} \frac{\partial \mathbf{Y}_{11}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{p1}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{Y}_{1q}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{pq}}{\partial x} \end{bmatrix} \in \mathbb{R}^{q \times p}$
$\mathbf{x} \in \mathbb{R}^m$	$\begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_m} \end{bmatrix} \in \mathbb{R}^m$	$\begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_p}{\partial \mathbf{x}_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_1}{\mathbf{x}_m} & \cdots & \frac{\mathbf{y}_p}{\partial \mathbf{x}_m} \end{bmatrix} \in \mathbb{R}^{m \times p}$	beyond this course
$\mathbf{X} \in \mathbb{R}^{m \times n}$	$\begin{bmatrix} \frac{\partial y}{\partial \mathbf{X}_{11}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial \mathbf{X}_{m1}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}$	beyond this course	beyond this course

Gradient Descent

- Gradient Descent is frequently used in Machine Learning algorithms, which means to use gradient to descent on the optimization hyper-plane. In one-variable and two-variable cases, the gradient descent look like



(a) One-variable Case **(b)** Two-variable Case

Figure. Gradient Descent Visualization Example

- Define the objective function you're about to optimize is $\mathcal{J}(\mathbf{w})$. The gradient on some value $\mathbf{w}^{(i)}$ is $\frac{\partial \mathcal{J}(\mathbf{w}^{(\mathbf{i})})}{\partial \mathbf{w}^{(\mathbf{i})}}$. The parameter update is $\mathbf{w}^{(\mathbf{i+1})} := \mathbf{w}^{(\mathbf{i})} - \alpha \frac{\partial \mathcal{J}(\mathbf{w}^{(\mathbf{i})})}{\partial \mathbf{w}^{(\mathbf{i})}}$, where the superscript (i) indicates the i-th times of optimization, α is called step size.

Code Implementation

- Closed-form Solution
- Gradient Descent
- Scikit-learn