Discussion #7

Dive into Gradient Descent

1. We want to minimize the loss function $L(\theta) = (\theta_1 - 1)^2 + |\theta_2 - 3|$. While you may notice that this function is not differentiable everywhere, we can still use gradient descent wherever the function *is* differentiable!

   Recall that for a function $f(x) = k|x|$, $\frac{df}{dx} = k$ for all $x > 0$ and $\frac{df}{dx} = -k$ for all $x < 0$.

   (a) What are the optimal values $\hat{\theta}_1$ and $\hat{\theta}_2$ to minimize $L(\theta)$? At that point $\hat{\theta}$, what is the gradient $\nabla L = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T \Big|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2}$?

   (b) Suppose we initialize our gradient descent algorithm randomly at $\theta_1 = 2$ and $\theta_2 = 5$. Calculate the gradient $\nabla L = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T \Big|_{\theta_1 = 2, \theta_2 = 5}$ at the specified θ_1 and θ_2 values.

(c) Apply the first gradient update with a learning rate $\alpha = 0.5$. In other words, calculate $\theta_1^{(1)}$ and $\theta_2^{(1)}$ using the initializations $\theta_1^{(0)} = 2$ and $\theta_2^{(0)} = 5$.

(d) How many gradient steps does it take for θ_1 and θ_2 to converge to their optimal values obtained in part (a) assuming we keep a constant learning rate of $\alpha = 0.5$? In other words, what is the value of t when $\theta^{(t)} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 \end{bmatrix}^T$.

Hint: After part (c), what is the derivative $\frac{\partial L}{\partial \theta_1}$ evaluated at $\theta_1^{(1)}$?

One-hot Encoding

2. In order to include a qualitative variable in a model, we convert it into a collection of Boolean vectors. These Boolean vectors contain only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 Boolean vectors that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 Boolean vectors for this dataset are x_A, x_B, and x_C, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{i,k} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let $\vec{y}$ represent any vector of outcome variables, and y_i is the value of said outcome for the i-th subject. This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, $\vec{x_C}$, and $\vec{y}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

(a) Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

(b) Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

(c) Show that

$$\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix}$$

where i is an element in group A, B, or C.

(d) Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned}
\hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\
&= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}
\end{aligned}$$

(e) *(Extra)* Show that if you augment your $\mathbb{X}$ matrix with an additional $\vec{1}$ bias vector as shown below, $\mathbb{X}^T\mathbb{X}$ is not full rank. Conclude that the new $\mathbb{X}^T\mathbb{X}$ is not invertible, and we cannot use the least squares estimate in this situation.

$$\mathbb{X} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | & | \end{bmatrix}$$