Linear Regression with Gradient Descent

Brian Pomerantz

October 2023

1 Linear Regression

The equation for linear regression is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b \tag{1}$$

where $\mathbf{X}$ is an $n \times f$ matrix where n is the number of samples and f is the number of features, $\hat{\mathbf{y}}$ is an $n \times 1$ column vector, $\mathbf{w}$ is an $f \times 1$ column vector, and b is a constant.

The bias term b can be folded into the weight vector $\mathbf{w}$ as an additional $f+1$ row by adding a column of 1s to $\mathbf{X}$, as shown below.

$$\mathbf{X} = \begin{pmatrix} x_{00} & x_{01} & ... & x_{0f} & 1 \\ x_{10} & x_{11} & ... & x_{1f} & 1 \\ ... \\ x_{n0} & x_{n1} & ... & w_{nf} & 1 \end{pmatrix} \tag{2}$$

and

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ ... \\ w_f \\ b \end{pmatrix} \tag{3}$$

Then equation (1) can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} \tag{4}$$

2 Loss Function

For the loss function, we use the Squared Loss function defined below.

$$\mathrm{SE}(\mathbf{w};\, \mathbf{X},\, \mathbf{y}) = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

where y_i is the i-th observation and $\hat{y}_i$ is the i-th row of $\hat{\mathbf{y}}$. In matrix form, equation (5) is equivalent to

$$\mathrm{SE}(\mathbf{w};\, \mathbf{X},\, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\top}(\mathbf{y} - \mathbf{X}\mathbf{w}) \tag{6}$$

where $\mathbf{y}$ is the $n \times 1$ column vector of observations. Note that the loss function is a function of the weight vector $\mathbf{w}$ with constant $\mathbf{X}$ and $\mathbf{y}$.

3 Gradient Descent

To perform gradient descent, we update our estimate of the weights $\mathbf{w}$ by moving in the direction which most quickly minimizes the loss function, *i.e.*,

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \gamma \frac{\partial \mathrm{SE}(\mathbf{w})}{\partial \mathbf{w}} \tag{7}$$

where γ is the learning rate and $\mathbf{w}_i$ is the i-th iteration of the gradient descent algorithm. With a suitable initial guess $\mathbf{w}_0$ and appropriate values of γ, convergence to a local minimum of the loss function is guaranteed so long as the loss function is convex and has a Lipschitz continuous gradient. Note that we have dropped the reference to constant $\mathbf{X}$ and $\mathbf{y}$ in equation (7).

Solving for the gradient of the loss function gives,

$$\begin{aligned} \frac{\partial \mathrm{SE}(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^{\top}(\mathbf{y} - \mathbf{X}\mathbf{w})\right] && (8) \\ &= \frac{\partial}{\partial \mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^{\top}\mathbf{I}_n(\mathbf{y} - \mathbf{X}\mathbf{w})\right] && (9) \end{aligned}$$

(where $\mathbf{I}_n$ is the $n \times n$ identity matrix) which according to equation (84) of *The Matrix Cookbook*,

$$\frac{\partial \mathrm{SE}(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^{\top}(\mathbf{y} - \mathbf{X}\mathbf{w}) \tag{10}$$

Therefore, combining equations (7) and (10), gradient descent for linear regression can be performed by updating the weight vector $\mathbf{w}$ for a given $\mathbf{X}$ and $\mathbf{y}$ according to the equation

$$\mathbf{w}_{i+1} = \mathbf{w}_i + 2\gamma\mathbf{X}^{\top}(\mathbf{y} - \mathbf{X}\mathbf{w}_i) \tag{11}$$