Linear Regression: Normal Equations[1]

Instead of using the gradient descent algorithm to iteratively minimize the sum of squared errors (SSE) criterion, we can minimize the SSE analytically and obtain a close form solution.

Given training data $(\mathbf{x}_i, y_i)$ for $i = 1, 2, ..., N$, with m input features, arranged in the design matrix X and the label vector $\mathbf{y}$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nm} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Then the error array associated with our regressor $f_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{m} w_k x_k$ is:

$$E = \left[\begin{array}{c} f_{\mathbf{w}}(\mathbf{x}_1) - y_1 \\ \dots \\ f_{\mathbf{w}}(\mathbf{x}_N) - y_N \end{array} \right] = \left[\begin{array}{c} \mathbf{x}_1^T \mathbf{w} \\ \dots \\ \mathbf{x}_N^T \mathbf{w} \end{array} \right] - \left[\begin{array}{c} y_1 \\ \dots \\ y_N \end{array} \right] = X\mathbf{w} - \mathbf{y}$$

(we used $w = (w_0, w_1, ...w_m)^T$ as a column vector). We can now write the sum of squares error as

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (f_w(\mathbf{x}_i) - y_i)^2 = \frac{1}{2} E^T E = \frac{1}{2} (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y})$$

Then,

$$\begin{aligned} \nabla_w J(\mathbf{w}) &= \nabla_w \frac{1}{2} E^T E = \nabla_w \frac{1}{2} (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y}) \\ &= \frac{1}{2} \nabla_w (\mathbf{w}^T X^T X\mathbf{w} - \mathbf{w}^T X^T \mathbf{y} - \mathbf{y}^T X\mathbf{w} + \mathbf{y}^T \mathbf{y}) \\ &= \frac{1}{2} \nabla_w (\mathbf{w}^T X^T X\mathbf{w} - \mathbf{w}^T X^T \mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) \\ &= \frac{1}{2} \nabla_w (\mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) \\ &= \frac{1}{2}\ 2(X^T X\mathbf{w} - X^T \mathbf{y}) \\ &= X^T X\mathbf{w} - X^T \mathbf{y} \end{aligned}$$

where, in the third step we have $(AB)^T = B^T A^T$.

[1] Based on lecture notes by Andrew Ng. These lecture notes are intended for in-class use only.

Since we are trying to minimize J, a convex function, a sure way to find w that minimizes J is to set its derivative to zero. In doing so we obtain

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

This is the exact $\mathbf{w}$ that minimizes the sum of squares error.