Linear Regression

- To find the optimal function $f(\mathbf{x}, \mathbf{w})$, we will pick that $\mathbf{w}$ which minimizes the error / loss between predictions $f(\mathbf{x}_i, \mathbf{w})$ and ground-truth labels y_i:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i, \mathbf{w}))^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

in the 1-dimensional case of $f: R \to R$.

- We set the derivatives wrt w_1, w_0 to zero to find the optimal w_1, w_0:

$$\frac{\partial L(\mathbf{w})}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial w_1} (y_i - w_0 - w_1 x_i)^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} 2(y_i - w_0 - w_1 x_i) \frac{\partial}{\partial w_1} (y_i - w_0 - w_1 x_i)$$

$$= \frac{2}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-x_i) = 0$$

$$\frac{\partial L(\mathbf{w})}{\partial w_0} = \frac{2}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i) \frac{\partial}{\partial w_0} (y_i - w_0 - w_1 x_i)$$

$$= \frac{2}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-1) = 0$$

($-1 = -x_0$, "feature" added for the bias)

- In multiple dimensions i.e. $f: R^D \to R$,

$$\frac{\partial L(\mathbf{w})}{\partial w_{j>0}} = \frac{2}{N} \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{D} w_j x_{ij}) \frac{\partial}{\partial w_j} (y_i - w_0 - \sum_{j=1}^{D} w_j x_{ij})$$

$$= \frac{2}{N} \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{D} w_j x_{ij})(-x_{ij})$$

where x_{ij} is the j-th dimension of the i-th sample.

– Using matrix notation, setting the derivative wrt $\mathbf{w}$ to 0:

Let $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & \cdots & x_1^D \\ \vdots & & & \\ 1 & x_N^1 & \cdots & x_N^D \end{bmatrix}$, $\mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix}$

Then $L(\mathbf{w}) = \frac{1}{N} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_1^1 & \cdots & x_1^D \\ 1 & x_N^1 & \cdots & x_N^D \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_D \end{bmatrix} \right\|^2$

$$= \frac{1}{N} \| \mathbf{y} - \mathbf{X}\mathbf{w} \|^2 = \frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

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

$$= \frac{1}{N} (\mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w})$$

$$\frac{\partial \mathbf{x}^T\mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{y} \qquad \frac{\partial \mathbf{x}^T\mathbf{y}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{y}\mathbf{x}$$

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N} [0 - \mathbf{X}^T\mathbf{y} - (\mathbf{y}^T\mathbf{X})^T + 2\mathbf{X}^T\mathbf{X}\mathbf{w}]$$

$$= \frac{1}{N} [-\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w}]$$

$$= -\frac{2}{N} [\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w}] = 0$$

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

$\mathbf{X}^\dagger$, Moore-Penrose pseudoinverse of $\mathbf{X}$

– To predict a label for a new sample $\mathbf{x}_{test}$:

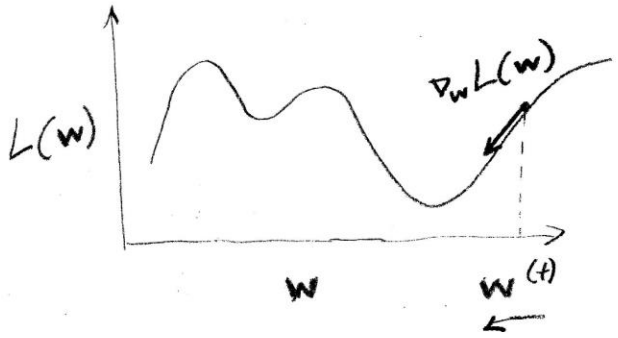
$$y = \mathbf{w}^{*T} \begin{bmatrix} 1 \\ \mathbf{x}_{test} \end{bmatrix} = (\mathbf{X}^\dagger\mathbf{y})^T \begin{bmatrix} 1 \\ \mathbf{x}_{test} \end{bmatrix} = \mathbf{y}^T \mathbf{X}^{\dagger T} \begin{bmatrix} 1 \\ \mathbf{x}_{test} \end{bmatrix}$$

Gradient Descent:

- Initialize $\mathbf{w}$ to a random vector.
- While $L(\mathbf{w})$ keeps changing, let $\nabla_{\mathbf{w}} L(\mathbf{w}) = \left[\frac{\partial L(\mathbf{w})}{\partial w_0} \quad \frac{\partial L(\mathbf{w})}{\partial w_1} \quad \cdots \quad \frac{\partial L(\mathbf{w})}{\partial w_D} \right]$, do:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}) \quad \text{(where } \eta \text{ is a "learning rate")}$$

$$\text{i.e. } w_j^{(t+1)} = w_j^{(t)} - \eta \frac{\partial L(\mathbf{w}^{(t)})}{\partial w_j}$$



Stochastic Gradient Descent:

- As above but use $L_i(\mathbf{w})$ defined for individual samples (randomly chosen).