Linear Regression

1 Hypothesis

Consider the domain $\mathcal{X} = \mathbb{R}^D$ and the codomain $\mathcal{Y} = \mathbb{R}$. In Linear Regression, we are trying to find an optimal hypothesis $h_\theta : \mathcal{X} \to \mathcal{Y}$ in the hypothesis class

$$\mathcal{H} = \left\{ h_\theta : h_\theta(\mathbf{x}) = \sum_{d=1}^{D} \theta^{(d)} x^{(d)} + \theta^{(0)}, \theta^{(i)} \in \mathbb{R} \right\}$$

which minimizes some loss function. Here, the loss function used will be the sum of squared errors loss, which is a very popular choice (the reason for which is detailed in **Bayesian Learning**). In other words, for each feature (dimension) of $\mathcal{X}$ we are finding an accompanying coefficient, and with those (and the offset $\theta^{(0)}$) we create a linear function that maps $\mathcal{X}$ to $\mathcal{Y}$. This can be visualized as a hyperplane in a standard graph.

2 Loss

Let the best hypothesis be denoted by h_{θ^*}. As mentioned, it is the one that minimizes the sum of squared errors under some given data. Let $\{\mathbf{x}_i\}_{i\in[N]}$ and $\{y_i\}_{i\in[N]}$ be the given data with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. We define a "squared error" simply by $(h_\theta(\mathbf{x}_i) - y_i)^2$. Thus, the sum of squared errors, our loss function, for some hypothesis h_θ is

$$L_S(h_\theta) = \frac{1}{2N} \sum_{i=1}^{N} (h_\theta(\mathbf{x}_i) - y_i)^2$$

The $\frac{1}{2N}$ term "normalizes" the error so that more data points don't increase the error (i.e. we are averaging the error). The 2 in the denominator serves to simplify the derivative of $L_S(h_\theta)$.

Let X be a $N \times (D+1)$ matrix with $\mathbf{x}$'s for rows prepended by 1's (this "trick" allows the offset terms to be introduced to the matrix formula presented later) and Y be a $N \times 1$ matrix with y's for entries, thus

$$X = \begin{bmatrix} 1 & - & \mathbf{x}_1 & - \\ 1 & - & \mathbf{x}_2 & - \\ & & \vdots & \\ 1 & - & \mathbf{x}_N & - \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

This allows us to rewrite the sum of squared errors loss function using matrix operations, thus

$$L_S(h_\theta) = \frac{1}{2N} (X\theta - Y)^\top (X\theta - Y)$$

We have that h_{θ^*} is the hypothesis that minimizes this loss, thus

$$h_{\theta^*} = \arg\min_{h_\theta} L_S(h_\theta)$$

Since L_S is quadratic in terms of θ (so it has a single global minima), this can be found through calculus, by finding the derivative of L_S and setting it to 0. Doing so, we obtain the closed form

$$\theta^* = (X^\top X)^{-1} X^\top Y$$

Other optimization methods are certainly possible, and information specifically pertaining to Linear Regression regarding them can be found in the following sections (although more detailed descriptions of these methods can be found in their respective documents).

3 Gradient Descent

Gradient descent is an efficient optimization option for finding θ^* since L_S is convex and its gradient is quite simple. It also preferable when X is too large (making $(X^\top X)^{-1}$ expensive to compute). The gradient of the loss function is

$$\nabla_\theta L_S(h_\theta) = \frac{1}{N} X^\top (X\theta - Y)$$

4 Stochastic Gradient Descent

We can use a random row in X as an unbiased estimate of the gradient (i.e. we can use some row $[1 \quad -\mathbf{x}_i-]$) for the SGD update step.

5 Caveats

If features are collinear, $(X^\top X)^{-1}$ will not be invertible, impeding the use of the closed form solution. L_S will also no longer have a single global minima, and so gradient descent may not converge.