Linear Regression with Basis Function Expansion

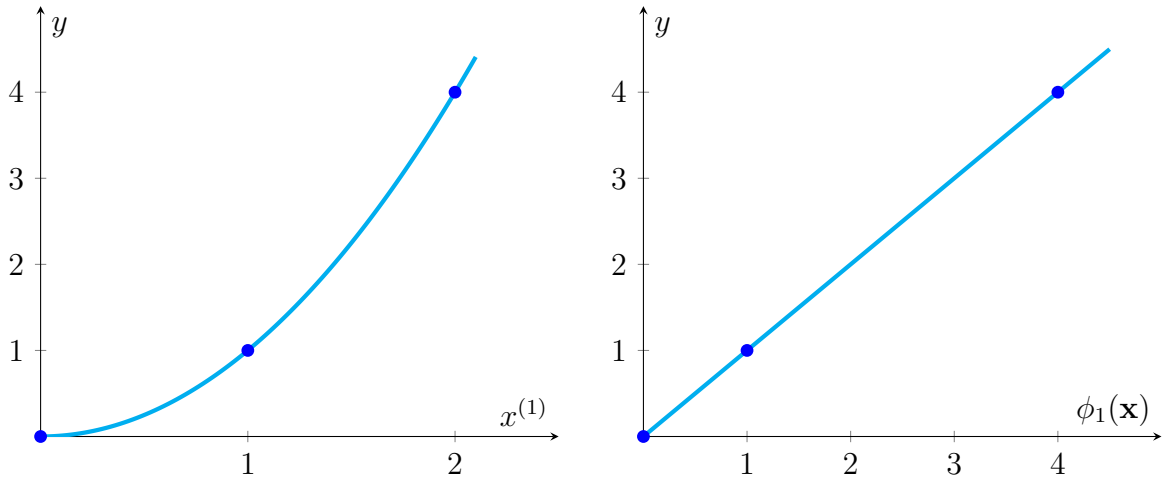
1 Hypothesis

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

$$\mathcal{H} = \left\{ h_\theta : h_\theta(\mathbf{x}) = \sum_{k=1}^{K} \theta^{(k)} \phi_k(\mathbf{x}) \right\}$$

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**). ϕ_i is some function such that $\phi_i : \mathcal{X} \to \mathbb{R}$ and K is how many of these functions we have. In other words, we are finding a linear combination of some set of transformations of the original features.

This can be thought of in two ways. In terms of the original feature space, this means we can find arbitrarily shaped functions to fit the data (limited only by our selection of ϕ's). In terms of the new feature space, created by all the transformations, we are still finding a hyperplane that best fits the new data points. This can be visualized in the example below, with $D = 1$, $K = 1$ and $\phi_1(\mathbf{x}) := (x^{(1)})^2$ (note that $\mathbf{x} = x^{(1)}$ in this case).



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 Φ be a $N \times K$ matrix with each row consisting of the transformations of the corresponding previous data point by each ϕ_i, and Y be a $N \times 1$ matrix with y's for entries, thus

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_K(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_K(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_K(\mathbf{x}_N) \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

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

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

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

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

However, due to the nature of the new hypothesis class, strictly minimizing the sum of squared errors may introduce overfitting. Tackling this issue is discussed in the next section.

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

3 Regularization

Due to overfitting, regularization can be applied by appending a term to the loss function that penalizes "large" θ's (which can be viewed as representing complex functions).

Further, regularization can make it possible to find a solution when Φ has more features than there are data points, which would cause L_S to not be strongly convex, leading to multiple global minima. If the regularization term is strongly convex, the new objective function becomes strongly convex, and so has one global minima.

As such, the optimal hypothesis can then be defined in terms of a new objective function $L_S(h_\theta) + R(h_\theta)$ like so

$$h_{\theta^*} = \arg\min_{h_\theta} \left(L_S(h_\theta) + R(h_\theta)\right)$$

where $L_S(h_\theta)$ is the previous loss function and $R(h_\theta)$ is the regularization term.

Different regularization techniques are described in the subsections below.

3.1 Ridge regression

Ridge regression or Tikhonov regularization utilizes the strongly convex regularization term $\frac{\lambda}{2N}\|\theta\|$. Thus, the new objective function is

$$\frac{1}{2N}(\Phi\theta - Y)^\top(\Phi\theta - Y) + \frac{\lambda}{2N}\|\theta\|^2$$

where λ controls the amount of regularization. The 2 in the denominator has the same purpose as before (simplifying the derivative of L_S). The closed form solution then becomes

$$\theta^* = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top Y$$

where I is the identity matrix.

3.2 LASSO

LASSO utilizes the non-strongly convex regularization term $\frac{\lambda}{N}\|\theta\|_1$. Thus, it is very similar to ridge regression but uses ℓ^1 instead of ℓ^2 norm. Using Lagrange multipliers, the minimization of the new objective function can also be written in terms of θ as

$$\arg\min_\theta \frac{1}{N}\sum_{i=1}^{N}(y_i - \Phi_i^\top\theta)^2, \quad \text{s.t.} \quad \sum_{i=1}^{K}\theta_i \le t$$

where Φ_i is a row of Φ and t is some constant corresponding to λ. Since the regularization term is not differentiable, a closed form solution cannot be derived.

4 Optimization

4.1 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 Φ is too large (making $(\Phi^\top\Phi)^{-1}$ expensive to compute). The gradient of the loss without regularization is

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

The gradient with ridge regression is

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

Since the regularization term in LASSO is not differentiable, a gradient cannot be computed.

5 Stochastic Gradient Descent

We can use a random row in Φ as an unbiased estimate of the gradient for the SGD update step.

6 Caveats

If features are collinear, $(\Phi^\top \Phi)^{-1}$ and $(\Phi^\top \Phi + N\lambda I)^{-1}$ 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.