Regression by linear combination of basis functions

Risi Kondor

February 5, 2004

Given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ where $x \in \mathcal{X}$ and $y \in \mathbb{R}$, the task of regression is to fit a real valued function $f : \mathcal{X} \mapsto \mathbb{R}$ to these points. In the simplest case $\mathcal{X} = \mathbb{R}$. In multidimensional regression $\mathcal{X} = \mathbb{R}^D$. It is sometimes necessary to do regression on more complicated spaces, but we are not going to deal with that here.

The easiest way to attack the regression problem is to look for f in a finite dimensional space of functions spanned by a given basis. In other words, we specify a set of functions $\phi_0, \phi_1, \ldots, \phi_P$ from $\mathcal{X}$ to $\mathbb{R}$ and look for f in the form of a linear combination

$$f(x) = \sum_{i=0}^{P} \theta_i \, \phi_i(x). \tag{1}$$

Performing the regression then reduces to finding the real parameters $\theta_1, \theta_2, \ldots, \theta_P$.

Different Bases

Linear regression

The simplest case is that of linear regression. In the one dimensional case we would simply take $\phi_0(x) = 1$ and $\phi_1(x) = x$. This gives

$$f(x) = \sum_{i=0}^{1} \theta_i \, \phi_i(x) = \theta_0 + \theta_1 x,$$

so by tuning θ_0 and θ_1 we can make f be any linear function. In the multidimensional case we would take $\phi_1(x) = [x]_1$, $\phi_2(x) = [x]_2$, all the way to $\phi_D(x) = [x]_D$. Here $[x]_i$ denotes the i'th component of the vector x. Unfortunately, we cannot use the simpler notation x_i for this purpose, because that is already reserved for the i'th data point. All linear functions $f : \mathbb{R}^D \mapsto \mathbb{R}$ can be expressed in this basis:

$$f(x) = \sum_{i=0}^{D} \theta_i \, \phi_i(x) = \theta_0 + \theta_1 \, [x]_1 + \theta_2 \, [x]_2 + \ldots + \theta_D \, [x]_D \, .$$

Realizing the constant term by setting $\phi_0(x) = 1$ will be common to all the function classes we discuss.

Polynomial regression

Another possible choice of basis (in the one-dimensional case) is to set $\phi_i(x) = x^i$ for $i = 1, 2, \ldots, P$. This lets us choose f from the class of polynomial functions of degree at most P:

$$f(x) = \sum_{i=0}^{P} \theta_i \, \phi_i(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_P x^P.$$

The multidimensional case is more complicated, because i has to become a multi-index $i = (i_1, i_2, \ldots, i_D)$ and the sum over i's becomes a sum over all multi-indices with $i_1 + i_2 + \ldots + i_d \leq P$. For example, for $D = 3$ and $P = 2$,

$$\begin{aligned} f(x) = \sum_{(i_1, i_2, i_3)} \theta_i \, \phi_i(x) = \\ \theta_0 + \theta_{(1,0,0)} \, [x]_1 + \theta_{(0,1,0)} \, [x]_2 + \theta_{(0,0,1)} \, [x]_3 + \theta_{(1,1,0)} \, [x]_1 \, [x]_2 + \\ \theta_{(0,1,1)} \, [x]_2 \, [x]_3 + \theta_{(1,0,1)} \, [x]_3 \, [x]_1 + \theta_{(2,0,0)} \, [x]_1^2 + \theta_{(0,2,0)} \, [x]_2^2 + \theta_{(0,0,2)} \, [x]_3^2 \end{aligned}$$

Gaussian RBF's

The last basis we look at is that of Gaussian Radial Basis Functions (RBF's)

$$\phi_z = e^{-\| x - z \|^2 / (2\sigma^2)}$$

where σ is a pre-set variance parameter. Of course, this would give an uncountably infinite number of basis functions (z can be anywhere in $\mathbb{R}^D$), which we cannot have. We remedy the situation by only considering Gaussian RBF's centered at the data points themselves,

$$\phi_i = e^{-\| x - x_i \|^2 / (2\sigma^2)}.$$

This step is not as arbitrary as it sounds, but we cannot descibe the justifiction for it here. As before, ϕ_0 is still the constant function $\phi_0(x) = 1$.

Solving for the θ's

To find the optimal value for $\theta_0, \theta_1, \ldots, \theta_P$ we

1. define a loss function L;
2. using the loss function define the empirical risk $R_{\text{emp}}(\theta)$ quantifying the loss over all the training data for particular values of $\theta_0, \theta_1, \ldots, \theta_P$;
3. solve for the particular setting of the parameters (denoted $\theta_0^*, \theta_1^*, \ldots, \theta_P^*$) that minimizes the empirical risk.

We shall use the squared error loss function

$$L(y, f(x)) = \frac{1}{2}\left(y - f(x)\right)^2 .$$

This is the simplest possible loss function, and it just says that the loss is proportional to the square of the difference between the predicted value and the true value. The empirical risk is then

$$R_{\text{emp}}(f) = R_{\text{emp}}(\theta_0, \theta_1, \ldots, \theta_P) = \frac{1}{2N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{P}\theta_j\,\phi_j(x_i)\right)^2 .$$

To simplify the development, we now introduce the vectors

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_P \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{pmatrix}$$

and the matrix

$$Q = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_P(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & & \phi_P(x_2) \\ \vdots & & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_P(x_N) \end{pmatrix} . \tag{2}$$

On the slides X is used for Q, but in the general case where ϕ_i are not linear functions that might be misleading. The empirical risk can then be written in the much shorter form

$$R_{\text{emp}} = \frac{1}{2N}\,\|\,\boldsymbol{y} - Q\boldsymbol{\theta}\,\|^2 .$$

To find $\boldsymbol{\theta}^*$, we can just set the derivatives of the empirical risk with respect to each θ_i equal to zero

$$\frac{\partial R_{\text{emp}}}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left[\frac{1}{2N}\,\|\,\boldsymbol{y} - Q\boldsymbol{\theta}\,\|^2\right] = 0$$

and solve for $\boldsymbol{\theta}$. In short hand, this is written as the single equation

$$\nabla_{\boldsymbol{\theta}} R_{\text{emp}} = 0.$$

We can then solve for the optimal $\boldsymbol{\theta}$ by

$$\begin{aligned}
0 &= \nabla_{\boldsymbol{\theta}} R_{\text{emp}} \\
&= \nabla_{\boldsymbol{\theta}} \left[\frac{1}{2N} \| \boldsymbol{y} - Q\boldsymbol{\theta} \|^2 \right] \\
&= \frac{1}{2N} \nabla_{\boldsymbol{\theta}} \left[(\boldsymbol{y} - Q\boldsymbol{\theta})^T (\boldsymbol{y} - Q\boldsymbol{\theta}) \right] \\
&= \frac{1}{2N} \nabla_{\boldsymbol{\theta}} \left[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T Q\boldsymbol{\theta} + \boldsymbol{\theta}^T Q^T Q \,\boldsymbol{\theta} \right] \\
&= \frac{1}{2N} \left(-2Q^T\boldsymbol{y} + 2Q^T Q \,\boldsymbol{\theta} \right)
\end{aligned}$$

leading to

$$Q^T Q \,\boldsymbol{\theta} = Q^T \boldsymbol{y}$$

and

$$\boldsymbol{\theta} = \left(Q^T Q \right)^{-1} Q^T \boldsymbol{y}. \tag{3}$$

This optimal value of $\boldsymbol{\theta}$ we denote $\boldsymbol{\theta}^*$.

In summary, we can use the same formula (3) no matter whether we do linear, polynomial or RBF regression, the only thing that changes is the definition of the matrix Q (2).

In fact, a notable simplification occurs in the RBF case, if we omit the bias term (leave $\phi_0(x) = 1$ out of the basis). In this case Q is a symmetric square basis, so (3) reduces to just $\boldsymbol{\theta}^* = Q^{-1}\boldsymbol{y}$.

To plot the resulting regression function we substite $\theta_1^*, \theta_1^*, \ldots, \theta_P^*$ back into (1). For example, in the RBF case

$$f(x) = \theta_0^* + \sum_{i=1}^{P} \theta_i^* \, e^{-\| x - x_i \|^2 / (2\sigma^2)}.$$