Regression Analysis

27 July 2012

Problem

We have a set of N data points $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. The idea is to come up with a best approximation fucntion that would predict the value of y for a given value of x. The nature of the underlying function that defines the data or the details of how the data was generated is usually unknown. These data points might be totally random or may be related by a characteristic function.

Theory

In the absence of the underlying function, the data is treated to be a linear combination of some set of functions.

$$\begin{aligned} y &= \sum_{i=1}^{M} w_i \phi_i(x) && (1) \\ &= \bar{w}^T \bar{\phi}(x) && (2) \end{aligned}$$

where M is the number of functions, $\bar{w} = [w_1 w_2 \ldots w_M]^T$, and $\bar{\phi}(x) = [\phi_1(x)\phi_2(x)\ldots\phi_M(x)]^T$. As a particular case, the set of functions $\{\phi_(x)\}$ belong to a set of orthogonal basis collection of functions. Any two functions in this orthogonal basis set satisy the following two criteria:-

$$\int_a^b \phi_i(x)\phi_j(x)dx = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

(1) is a *linear model* for regression. Approximating the output values y_n using the linear model results in an error. Minimizing this error forms the backbone of this approach. The error in approximation is given by $(y_n - \hat{y_n})^2$ where $\hat{y_n}$ is the estimated y value. The combined error for all N data points can then be written as

$$\mathcal{E} = \sum_{n=1}^{N} (y_n - \hat{y_n})^2 \qquad (3)$$

The error given a M and $\bar{w}$ is

$$\mathcal{E}(\bar{w}) = \sum_{n=1}^{N} (y_n - \sum_{i=1}^{M} w_i \phi_i(x_n))^2 \qquad (4)$$

The optimal set of weights are those that minimize $\mathcal{E}(\bar{w})$

$$\bar{w}^* = \underset{\bar{w}}{\operatorname{argmin}}\, \mathcal{E}(\bar{w})$$

Differentiating $\mathcal{E}(\bar{w})$ with respect to $\bar{w}$, we have

$$\begin{aligned}\frac{d}{d\bar{w}}\mathcal{E}(\bar{w}) &= \frac{d}{d\bar{w}}\sum_{n=1}^{N}(y_n - \bar{w}^T\bar{\phi}(x_n))^2 \\ &= 2\sum_{n=1}^{N}(y_n - \bar{\phi}(x_n)^T\bar{w})\bar{\phi}(x_n)\end{aligned}$$

Now,

$$\frac{d}{d\bar{w}}\mathcal{E}(\bar{w}) = 0 \Rightarrow \sum_{n=1}^{N}(y_n - \bar{\phi}(x_n)^T\bar{w})\bar{\phi}(x_n) = 0$$

$$\therefore \sum_{n=1}^{N} y_n\bar{\phi}(x_n) = \sum_{n=1}^{N}\bar{\phi}(x_n)^T\bar{w}\bar{\phi}(x_n) \tag{5}$$

If $\bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix}$ and $\Phi = \begin{bmatrix} \bar{\phi}(x_1)^T \\ \bar{\phi}(x_2)^T \\ \cdot \\ \cdot \\ \cdot \\ \bar{\phi}(x_N)^T \end{bmatrix}_{N\times M}$, then (5) can be expressed as

$$\begin{aligned}(\Phi^T\Phi)\bar{w} &= \Phi^T\bar{y} \\ \bar{w} &= (\Phi^T\Phi)^{-1}\Phi^T\bar{y} \qquad (6)\end{aligned}$$

Experiment

Tests are done by simulating data from two functions in particular, namely, the `sawtooth` and `square` functions. Each of these functions can be represented using an infinite Fourier series representation.

Data Generation

- *Generating X's:* The range from which the x values need to be generated is defined at runtime via command-line arguments using the parameters `low` and `high`. Number of samples (`nsamples`) is also specified at runtime. Using a random data generator, these x values are obtained.

- Corresponding to a particular function that is also specified at runtime, the function values $f(x)$ for the respective x's are computed.

- *Generating Y's:* To the previously generated $f(x)$ values, some amount of *Gaussian noise* is added to account for any errors in the actual experiment conducted.

$$y = f(x) + \epsilon \quad \text{and} \quad \epsilon \sim \mathcal{N}(\mu, \sigma)$$