Regression Analysis

1 Oct 2012

(I) Problem

The problem is as follows: Given a set of N data points

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\},$$

find a function that best approximates the underlying function $f(x_i) = y_i$ which relates the data. 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.

(II) Theory

In the absence of the underlying function, one could treat the data as a linear combination of some set of functions.

$$y = \sum_{i=1}^{M} w_i \phi_i(x) \tag{1}$$

$$= \bar{w}^T \bar{\phi}(x), \tag{2}$$

where M is the number of terms, $\bar{w} = [w_1 w_2 \ldots w_M]^T$ (superscript T refers to the matrix transpose), and $\bar{\phi}(x) = [\phi_1(x)\phi_2(x) \ldots \phi_M(x)]^T$. Here, w_i is the weight corresponding to the function ϕ_i.

II.I Orthogonal Functions

Consider two functions $\phi_1(x)$ and $\phi_2(x)$ defined over the range $[a, b]$. The *inner product* of these functions is defined as

$$< \phi_1, \phi_2 >= \int_a^b \phi_1(x)\phi_2(x)$$

ϕ_1 and ϕ_2 are said to be *orthogonal* if $< \phi_1, \phi_2 >= 0$. If there is a set of functions $\{\phi_1, \phi_2, \ldots, \phi_M\}$ defined over a range, then these functions form an orthogonal set if the inner product of any pair of these orthogonal functions $< \phi_i, \phi_j >= 0$ for $i \neq j$.

II.II Fourier Series

A Fourier series is a decomposition of a periodic function into a sum of

infinite sine/cosine functions. Any periodic function $f(x)$ with fundamental period T can be represented using Fourier series as:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n cos\frac{2n\pi x}{T} + b_n sin\frac{2n\pi x}{T} \right)$$

The Fourier coefficients a_n and b_n can be determined from the following integrals:

$$a_0 = \frac{2}{T}\int_0^T f(x)dx \tag{3}$$

$$a_n = \frac{2}{T}\int_0^T f(x)cos\frac{2n\pi x}{T}dx \tag{4}$$

$$b_n = \frac{2}{T}\int_0^T f(x)sin\frac{2n\pi x}{T}dx \tag{5}$$

(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

$$\mathcal{E} = \sum_{n=1}^{N}(y_n - \hat{y_n})^2 \tag{6}$$

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

$$\mathcal{E}(\bar{w}) = \sum_{n=1}^{N}(y_n - \sum_{i=1}^{M} w_i\phi_i(x_n))^2 \tag{7}$$

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,

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

$$\therefore \sum_{n=1}^{N} y_n\bar{\phi}(x_n) = \sum_{n=1}^{N}\bar{\phi}(x_n)^T\bar{w}\bar{\phi}(x_n) \tag{8}$$

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

$$(\Phi^T \Phi)\bar{w} = \Phi^T \bar{y}$$
$$\bar{w} = (\Phi^T \Phi)^{-1} \Phi^T \bar{y} \tag{9}$$

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.

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