Why

We use a loss function and the predictive density of some dataset to construct a regressor on the domain of the random process.

Setup

Let $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be a loss function. We choose a predictor to minimize the expected loss under the predictive density.

Squared Error Case

Consider $\ell(\alpha, \beta) = (\alpha - \beta)^2$. The *minimum squared error normal random function predictor* or *minimum squared error gaussian process predictor* for dataset $(a^1, \gamma_1), \ldots, (a^n, \gamma_n)$ in $A \times \mathbf{R}$ is the predictor which minimizes the squared error loss. Since the predictive density is normal, the minimizer is the conditional mean.

Absolute Error Case

Consider $\ell(\alpha, \beta) = |\alpha - \beta|$. The *minimum absolute deviation normal random function predictor* or *minimum absolute deviation gaussian process predictor* for dataset $(a^1, \gamma_1), \ldots, (a^n, \gamma_n)$ in $A \times \mathbf{R}$ is the predictor which minimizes the absolute deviation loss. For any density, the solution is the median. Since the predictive density is normal, and so symmetric, the median is the conditional mean. In other words, the minimum absolute deviation normal random function predictor coinicides with

the minimium squared error normal random function predictor.

Definition

For this reason, the *normal random function predictor* or *gaussian process predictor* for dataset $(a^1, \gamma_1), \ldots, (a^n, \gamma_n)$ in $A \times \mathbf{R}$ is $h : A \to \mathbf{R}$ defined by

$$h(x) = m(x) + \left(k(x, a^1) \cdots k(x, a^n) \right) (\Sigma_a + \Sigma_e)^{-1} (\gamma - m_a).$$

In other words, the regressor which assigns to each point its conditional mean.

Notice that h is an affine function of γ. If the mean function $m \equiv 0$ then h is linear in γ. This is sometimes called a *linear estimator*.[1] Alternatively, notice (in the zero mean setting) that h is a linear combination of n kernel function $k(x, a^i)$ for $i = 1, \ldots, n$. Specifically, h is a linear combination of

The process of using a normal random function predictor is often called *Gaussian process regression* or (especially in spatial statistics) *kriging*. The upside is that a gaussian process predicor interpolates the data, is smooth, and the so-called variance increases with the distance from the data.[2]

---

[1] We avoid the other terminology we have seen used—linear predictor—because the predictor h is not linear in its input x.

[2] This last piece is true for certain covariance kernels and will be clarified in future editions.