Bayesian Gaussian Processes

Casey Davis

This Github Repository

This particular Github repo, *bayesianGPs*, is a rough draft of putting together some software that will do the stuff in my dissertation, *A Bayesian Approach to Prediction and Variable Selection Using Non-Stationary Gaussian Processes*. This one will use Stan, a Bayesian programming language/software. I might also duplicate in R using Rcpp what I did when I originally wrote this software in MATLAB, but that'll probably only be to practice certain things and random experimentation.

I'll probably build up from the most basic stationary model and work towards the stuff in my dissertation. Eventually, I'll make an actual R package (Python, too?) in a different repo.

Setup

We have observed data, $\mathbf{Y}$. Conditional on model parameters, Λ, the data are treated as a realization from a Gaussian process, $Y(\mathbf{x})$.

$$Y(\mathbf{x})|\Lambda \sim GP\left(\mu, C\left(\cdot, \cdot\right)\right)$$

where μ is the process mean, and $C\left(\cdot, \cdot\right)$ is the process covariance.

$$C\left(\mathbf{x}, \mathbf{x}'\right) = Cov\left(Y\left(\mathbf{x}\right), Y\left(\mathbf{x}'\right)\right)$$

A common covariance function is the stationary *squared exponential* or *Gaussian* covariance function. For $\mathbf{x} \in \mathbb{R}^D$, it commonly is seen in some version of the form

$$C\left(\mathbf{x}, \mathbf{x}'\right) = \sigma^2 exp\left(-\sum_{d=1}^{D}\theta_d\left(x_d - x_d'\right)^2\right) + \sigma_\epsilon^2\delta$$

where σ^2 is the process variance, θ_d is a scale-parameter for each dimension d, σ_ϵ^2 is the scale of random noise, and δ is the Kronecker delta function with $\delta = 1$ if $\mathbf{x} = \mathbf{x}'$ and 0 otherwise. I prefer a slightly different parameterization that eases interpretability:

$$C\left(\mathbf{x}, \mathbf{x}'\right) = \sigma^2 \prod_{d=1}^{D} \rho_d^{16\left(x_d - x_d'\right)^2}$$

In this parameterization, ρ_d is the correlation between the response at two points that differ only in the dth dimension by $\frac{1}{4}$ of a unit. So if the inputs are all standardized to $[0, 1]^D$, then ρ_d is the correlation between the response at two points that differ only in the dth dimension by $\frac{1}{4}$ the range of the data in the dth dimension.

Model

We have n observations, $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ at n locations, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})^\top$, where $\mathbf{x_i} = (x_{i,1}, \ldots, x_{i,d})^\top$. Basically, the ith row of $\mathbf{X}$ contains the input values for the ith observation.

There are two formulations for Gaussian data. First is the latent variable GP:

$$y_i|\mathbf{\Lambda}, f_i \sim N(f_i, \sigma_\epsilon^2), i = 1, \ldots, n$$

$$\mathbf{f}|\mathbf{\Lambda} \sim N_n\left(\mu\mathbf{1_n}, \mathbf{K}\right)$$

$$\rho_d \sim Beta\left(\alpha_d, \beta_d\right), d = 1, \ldots, D$$

$$\sigma \sim whatever$$

$$\sigma_\epsilon \sim whatever$$

$$p(\mu) \propto 1$$