Summary of "Warped Gaussian Processes"

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 3 March, 2014

Snelson, Rasmussen, and Ghahramani (2003) generalize Gaussian processes (GP) to include a nonlinear transformation step. The advantage of this generalization is that the transformation is chosen algorithmically so that the transformed data is well-modeled by a GP. This stands in contrast to ad hoc procedures used by researchers to coerce their data into a GP model, such as the log transformation. The authors refer to their approach as a warped Gaussian process (wGP).

The wGP is intended for regression applicaitons, in which the researcher seeks the predictive distribution $P(t_{N+1}|x^{n+1}, \mathcal{D})$. Because the GP prior gives rise to a multivariate Gaussian distribution over functions y, we need a covariance function $C(x, x')$. Adding the warping step, parameterized by Psi, transforms the negative log likelihood from the usual

$$\begin{aligned}\mathcal{L} &= -\log P(t_N|X_N, \Theta) \\ &= \frac{1}{2}\log \det C_N + \frac{1}{2}t_n^\top C_N^{-1} t_n + \frac{N}{2}\log 2\pi\end{aligned}$$

to

$$\begin{aligned}\mathcal{L} &= -\log P(t_N|X_N, \Theta, \Psi) \\ &= \frac{1}{2}\log \det C_N + \frac{1}{2}f(t_n)^\top C_N^{-1} f(t_n) - \sum_{n=1}^{N} \log \frac{\partial f(t)}{\partial t}\Big|_{t_n} + \frac{N}{2}\log 2\pi,\end{aligned}$$

which basically replaces the t_n term with $f(t_n|\Psi)$ and the Jacobian to incorporate the transformation $f(\cdot)$. For prediction we also need the inverse function f^{-1}, but if it is unavailable analytically (as is often the case) it can be approximated with Newton-Raphson.

When is the warped GP preferable to the standard approach? One situation that would give rise to the wGP is when the values of the inputs are associated with bias in the outputs, especially if the bias increased with the value of the inputs. If the data is already modeled well by the a GP, the transformation f will be linear and can be avoided entirely. In some cases the log transformation is also sufficient; there are diminishing returns to using the warped GP in many real datasets (see Table 1 in the paper). The warped GP tends to perform poorly when the data has been censored at some threshold so that many data points lie immediately above or below the cut-off. In such a case the wGP will increase the variance in the transformed points, whereas a censorship model would be more principled. Overall the wGP offers a more principled approach for transforming inputs, but adds a number of "researcher degrees of freedom" by allowing the modeler to control the size, steepness, position, and number of steps for determining the function f.