Bayesian Regression using Gaussian Process priors

Carl Edward Rasmussen
Dept. of Computer Science
University of Toronto
carl@cs.toronto.edu
http://www.cs.toronto.edu/~carl

Outline

- The theory of non-parametric regression with Gaussian Process priors
- Applying the model to data
- Computational considerations

The work presented here has been done in collaboration with Chris Williams (Aston) and Radford Neal (Toronto).

Similar models were explored 20 years ago by O'Hagan [1978] but have not attracted much attension.

The task

Given n noisy training cases

$$\mathcal{D} = \Big\{ \Big(x^{(i)}, t^{(i)}\Big) \Big| i = 1 \ldots n \Big\}, \tag{1}$$

and a novel input $x^{(n+1)}$, guess the output $y^{(n+1)}$.

$x^{(i)} = \Big(x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}\Big)$ are input vectors (predictors) of dimension m.

$t^{(i)}$ are the observed noisy scalar targets (responses).
$y^{(i)}$ are model outputs; it models the expected value of the targets.

Example — Linear model:

$$y^{(i)} = x^{(i)}\beta, \qquad t^{(i)} = y^{(i)} + \varepsilon. \tag{2}$$

In our Bayesian model, both y's and t's are random variables.

Formally, compute

$$p\Big(y^{(n+1)} \Big| x^{(n+1)}, \mathcal{D}\Big). \tag{3}$$

Requisites

Gaussian Process: any finite number of variables, Y_x, have a joint multivariate Gaussian distribution; I call these variables $y^{(i)}$.

Gaussian Process Prior over functions

$$p\Big(y^{(1)}, \ldots, y^{(n+1)} \Big| x^{(1)}, \ldots, x^{(n+1)}\Big) \propto \exp\Big(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\Big). \tag{4}$$

Example covariance function

$$\begin{aligned} \Sigma_{pq} &= \mathrm{cov}\Big(y^{(p)}, y^{(q)}\Big) = C\Big(x^{(p)}, x^{(q)}\Big) \\ &\propto \exp\Big(-\frac{1}{2}\sum_{j=1}^{m}\Big(x_j^{(p)} - x_j^{(q)}\Big)^2\Big). \end{aligned} \tag{5}$$

Likelihood: assume Gaussian noise:

$$p\Big(t^{(1)}, \ldots, t^{(n)} \Big| y^{(1)}, \ldots, y^{(n)}\Big) \propto \exp\Big(-\frac{1}{2}(\mathbf{y} - \mathbf{t})^\top \Omega^{-1} (\mathbf{y} - \mathbf{t})\Big), \tag{6}$$

where $\Omega = r^2 I$.

Making inference

Posterior, by Bayes' rule

$$\begin{aligned} &p\Big(y^{(1)}, \ldots, y^{(n+1)} \Big| x^{(n+1)}, \mathcal{D}\Big) \propto \\ &p\Big(t^{(1)}, \ldots, t^{(n)}, y^{(1)}, \ldots, y^{(n+1)} \Big| x^{(1)}, \ldots, x^{(n+1)}\Big). \end{aligned} \tag{7}$$

Perform two steps on the posterior distribution:

1) Marginalise $y^{(1)}, \ldots, y^{(n)}$ to obtain

$$p\Big(t^{(1)}, \ldots, t^{(n)}, y^{(n+1)} \Big| x^{(1)}, \ldots, x^{(n+1)}\Big) \propto \exp\Big(-\frac{1}{2}\tilde{\mathbf{t}}^\top \Big(\Sigma^{-1} + \Omega^{-1}\Big)\tilde{\mathbf{t}}\Big), \tag{8}$$

where $\tilde{\mathbf{t}} = \Big(t^{(1)}, \ldots, t^{(n)}, y^{(n+1)}\Big)$.

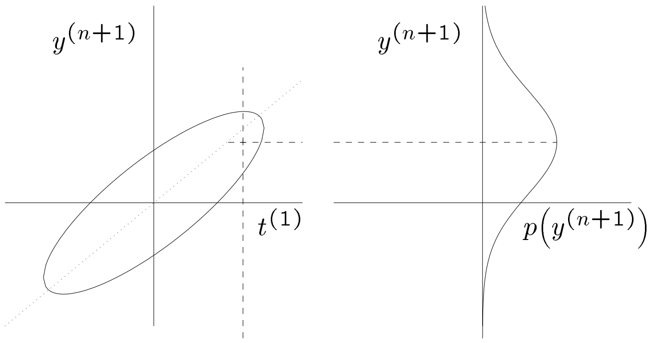
2) Condition on the observed targets to obtain

$$p\Big(y^{(n+1)} \Big| x^{(n+1)}, \mathcal{D}\Big) \sim \mathcal{N}(\mu, \sigma^2), \tag{9}$$

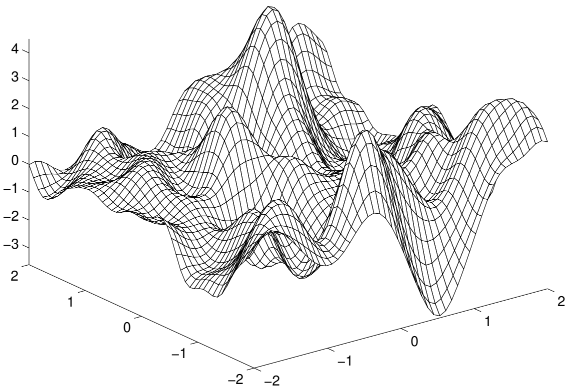
where

$$\begin{aligned} &\mu = \mathbf{a}^\top Q^{-1}\mathbf{t} \qquad \sigma^2 = b - \mathbf{a}^\top Q^{-1}\mathbf{a} \\ &Q_{pq} = C\Big(x^{(p)}, x^{(q)}\Big) + r^2\delta_{pq}, \quad p, q = 1, \ldots, n \\ &\mathbf{a}_p = C\Big(x^{(n+1)}, x^{(p)}\Big), \quad p = 1, \ldots, n \\ &b = C\Big(x^{(n+1)}, x^{(n+1)}\Big). \end{aligned} \tag{10}$$

Illustrations



Conditioning on the observed target in step 2 when $n = 1$.



Function drawn at random from a 2-d Gaussian Process prior.

Applying the model to data

Using a parameterised *Automatic Relevance Determination* (ARD) covariance function

$$C\Big(x^{(p)}, x^{(q)}\Big) = \nu \exp\Big(-\frac{1}{2}\sum_{j=1}^{m} w_j\Big(x_j^{(p)} - x_j^{(q)}\Big)^2\Big), \quad (11)$$

with *hyperparameters* $\theta = (\nu, w_1, \ldots, w_m, r^2)$.

The likelihood for θ is

$$\begin{aligned} &\log p(\mathcal{D}|\theta) \\ &\quad = \log p\Big(t^{(1)}, \ldots, t^{(n)} \Big| x^{(1)}, \ldots, x^{(n)}, \theta\Big) \\ &\quad = -\frac{1}{2}\log\det Q - \frac{1}{2}\mathbf{t}^\top Q^{-1}\mathbf{t} - \frac{n}{2}\log(2\pi), \end{aligned} \quad (12)$$

with partial derivatives

$$\begin{aligned} &\frac{\partial}{\partial\theta_i}\log p(\mathcal{D}|\theta) \\ &\quad = -\frac{1}{2}\mathrm{trace}\Big(Q^{-1}\frac{\partial Q}{\partial\theta_i}\Big) + \frac{1}{2}\mathbf{t}^\top Q^{-1}\frac{\partial Q}{\partial\theta_i}Q^{-1}\mathbf{t} \end{aligned} \quad (13)$$

Two possible approaches, gradient based ML (or MAP) or attempt to integrate out θ via MCMC.

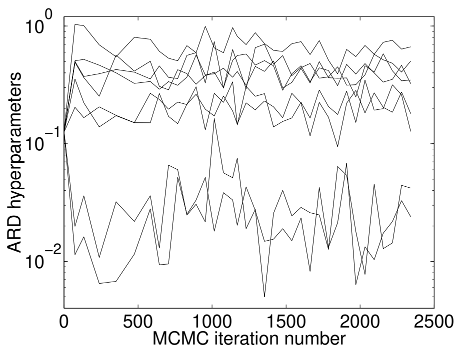
Predictions using MCMC

Set (vague) priors on the hyperparameters; approximate the predictive distribution

$$\begin{aligned} &p\Big(y^{(n+1)} \Big| x^{(n+1)}, \mathcal{D}\Big) \\ &\qquad = \int p\Big(y^{(n+1)} \Big| x^{(n+1)}, \theta\Big) p(\theta|\mathcal{D}) d\theta \\ &\qquad \simeq \frac{1}{T}\sum_{\tau=1}^{T} p\Big(y^{(n+1)} \Big| x^{(n+1)}, \theta^{(\tau)}\Big), \end{aligned} \quad (14)$$

where $\theta^{(\tau)}$ are samples from the posterior distribution for θ. The predictive distribution is a mixture of Gaussians.

The *Hybrid* Monte Carlo algorithm [Duane et al. 1987] has been used to avoid random walks in the exploration of θ.



Conclusions and Future directions

Simple non-parametric model; excellent predictive performance (comp. to linear models, kNN, MARS, NN, and Bayesian NN).
Gives predictive distribution
Handles several hundred training cases
Interpretability (ARD)

Additive models and linear contributions
Extension to classification
Extension to other noise models
Fast methods for approximate matrix inversions

Duane, S., Kennedy, A.D., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters, B*, 195:612-222.

O'Hagan, A. (1978). On curve fitting and optimal design for regression. *J.R.S.S., B*, 40:1-42.

Williams, C.K.I. and Rasmussen, C.E. (1996). Regression with Gaussian Processes, *Advances in Neural Information Processing Systems 8*, 290-297, MIT Press.