Linear in the parameters models and GP

Carl Edward Rasmussen

October 13th, 2016

Key concepts

- We give an interpretation of the marginal likelihood in terms of
  - a data fit
  - a complexity penalty
- covariance functions can be parameterized using hyperparameters
- hyperparameters can be fit by optimizing the marginal likelihood
  - this is a form of model selection
- Occam's razor is automatic and avoids overfitting

From random functions to covariance functions

Consider the class of linear functions:

$$f(x) \;=\; ax + b, \quad \text{where} \;\; a \sim \mathcal{N}(0, \alpha), \quad \text{and} \;\; b \sim \mathcal{N}(0, \beta).$$

We can compute the mean function:

$$\mu(x) \;=\; E[f(x)] \;=\; \iint f(x)p(a)p(b)\,da\,db \;=\; \int axp(a)\,da + \int bp(b)\,db = 0,$$

and covariance function:

$$\begin{aligned} k(x, x') \;&= E[(f(x) - 0)(f(x') - 0)] \;=\; \iint (ax + b)(ax' + b)p(a)p(b)\,da\,db \\ &= \int a^2 xx' p(a)\,da + \int b^2 p(b)\,db + (x + x') \int abp(a)p(b)\,da\,db = \alpha xx' + \beta. \end{aligned}$$

From finite linear models to Gaussian processes (1)

Finite linear model with Gaussian priors on the weights:

$$f(x) \;=\; \sum_{m=1}^{M} w_m \,\phi_m(x) \qquad\qquad p(\mathbf{w}) = \mathcal{N}(\mathbf{w};\; \mathbf{0}, A)$$

The joint distribution of any $\mathbf{f} = [f(x_1), \ldots, f(x_N)]^\top$ is a multivariate Gaussian – this looks like a Gaussian Process!
The prior $p(\mathbf{f})$ is fully characterized by the *mean* and *covariance* functions.

$$m(x) \;=\; E_{\mathbf{w}}\big(f(x)\big) \;=\; \int \Big(\sum_{m=1}^{M} w_k \phi_m(x)\Big) p(\mathbf{w}) d\mathbf{w} \;=\; \sum_{m=1}^{M} \phi_m(x) \int w_m p(\mathbf{w}) d\mathbf{w}$$

$$= \sum_{m=1}^{M} \phi_m(x) \int w_m p(w_m) dw_m \;=\; 0$$

The *mean function* is zero.

From finite linear models to Gaussian processes (2)

Covariance function of a finite linear model

$$f(x) = \sum_{m=1}^{M} w_m\,\phi_m(x) = \mathbf{w}^\top \boldsymbol{\phi}(x) \qquad \boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_M(x)]^\top{}_{(M\times 1)}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w};\, \mathbf{0}, A)$$

$$k(x_i, x_j) = \mathrm{Cov}_{\mathbf{w}}\big(f(x_i), f(x_j)\big) = E_{\mathbf{w}}\big(f(x_i)f(x_j)\big) - \underbrace{E_{\mathbf{w}}\big(f(x_i)\big)E_{\mathbf{w}}\big(f(x_j)\big)}_{0}$$

$$= \int\!\!\ldots\!\!\int \Big(\sum_{k=1}^{M}\sum_{l=1}^{M} w_k w_l \phi_k(x_i)\phi_l(x_j)\Big) p(\mathbf{w})\, d\mathbf{w}$$

$$= \sum_{k=1}^{M}\sum_{l=1}^{M} \phi_k(x_i)\phi_l(x_j) \underbrace{\iint w_k w_l\, p(w_k, w_l)\, dw_k\, dw_l}_{A_{kl}} = \sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\phi_k(x_i)\phi_l(x_j)$$

$$\boxed{k(x_i, x_j) = \boldsymbol{\phi}(x_i)^\top A \boldsymbol{\phi}(x_j)}$$

Note: If $A = \sigma_{\mathbf{w}}^2 \mathbf{I}$ then $k(x_i, x_j) = \sigma_{\mathbf{w}}^2 \sum_{k=1}^{M} \phi_k(x_i)\phi_k(x_j) = \sigma_{\mathbf{w}}^2 \boldsymbol{\phi}(x_i)^\top \boldsymbol{\phi}(x_j)$

GPs and Linear in the parameters models are equivalent

We've seen that a Linear in the parameters model, with a Gaussian prior on the weights is also a GP.

Note the different computational complexity: GP: $\mathcal{O}(N^3)$, linear model $\mathcal{O}(NM^2)$ where M is the number of basis functions and N the number of training cases.

So, which representation is most efficient?

Might it also be the case that every GP corresponds to a Linear in the parameters model? (Mercer's theorem.)