Bayesian inference and prediction with finite regression models

Carl Edward Rasmussen

July 1st, 2016

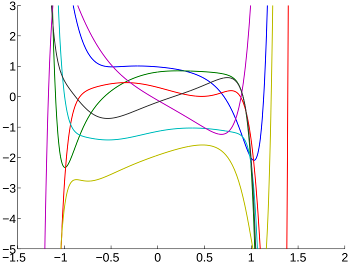
Key concepts

Posterior probability of a function

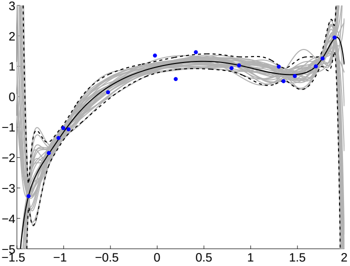
Given the prior functions $p(\mathbf{f})$ how can we make predictions?

- Of all functions generated from the prior, keep those that fit the data.
- The notion of closeness to the data is given by the likelihood $p(\mathbf{y}|\mathbf{f})$.
- We are really interested in the posterior distribution over functions:

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})\,p(\mathbf{f})}{p(\mathbf{y})} \qquad \text{Bayes Rule}$$



Some samples from the prior



Samples from the posterior

Priors on parameters induce priors on functions

A model $\mathcal{M}$ is the choice of a model structure and of para meter values.

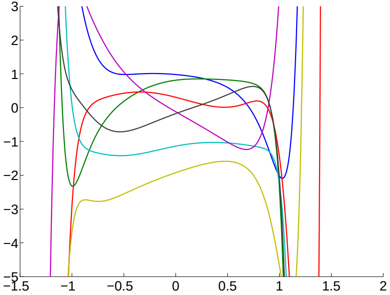
$$f_{\mathbf{w}}(x) \;=\; \sum_{m=0}^{M} w_m\, \phi_m(x)$$

The prior $p(\mathbf{w}|\mathcal{M})$ determines what functions this model can generate. Example:

- Imagine we choose $M = 17$, and $p(w_m) = \mathcal{N}(w_m;\ 0, \sigma_{\mathbf{w}}^2)$.
- We have actually defined a prior distribution over functions $p(\mathbf{f}|\mathcal{M})$.

This figure is generated as follows:

- Use polynomial basis functions, $\phi_m(x) = x^m$.
- Define a uniform grid of $n = 100$ values in x from $[-1.5, 2]$.
- Generate matrix $\mathbf{\Phi}$ for $M = 17$.
- Draw $w_m \sim \mathcal{N}(0, 1)$.
- Compute and plot $\mathbf{f} = \mathbf{\Phi}_{n\times 18}\,\mathbf{w}$.



Maximum likelihood, parametric model

Supervised parametric learning:

- data: $\mathbf{x}, \mathbf{y}$
- model $\mathcal{M}$: $y = f_{\mathbf{w}}(x) + \varepsilon$

Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathcal{M}) \;\propto\; \prod_{n=1}^{N} \exp(-\tfrac{1}{2}(y_n - f_{\mathbf{w}}(x_n))^2/\sigma^2_{\text{noise}}).$$

Maximize the likelihood:

$$\mathbf{w}_{\text{ML}} \;=\; \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathcal{M}).$$

Make predictions, by plugging in the ML estimate:

$$p(y_*|x_*, \mathbf{w}_{\text{ML}}, \mathcal{M})$$

Bayesian inference, parametric model, cont.

Posterior parameter distribution by Bayes rule ($p(a|b) = p(a)p(b|a)/p(b)$):

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \mathcal{M}) \;=\; \frac{p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathcal{M})}{p(\mathbf{y}|\mathbf{x}, \mathcal{M})}$$

Making predictions (marginalizing out the parameters):

$$\begin{aligned} p(y_*|x_*, \mathbf{x}, \mathbf{y}, \mathcal{M}) \;&=\; \int p(y_*, \mathbf{w}|\mathbf{x}, \mathbf{y}, x_*, \mathcal{M})d\mathbf{w} \\ &=\; \int p(y_*|\mathbf{w}, x_*, \mathcal{M})p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \mathcal{M})d\mathbf{w}. \end{aligned}$$

Posterior and predictive distribution in detail

For a linear-in-the-parameters model with Gaussian priors and Gaussian noise:

- Gaussian *prior* on the weights: $p(\mathbf{w}|\mathcal{M}) = \mathcal{N}(\mathbf{w};\ \mathbf{0},\ \sigma_{\mathbf{w}}^2\,\mathbf{I})$
- Gaussian *likelihood* of the weights: $p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M}) = \mathcal{N}(\mathbf{y};\ \mathbf{\Phi}\,\mathbf{w},\ \sigma_{\text{noise}}^2\,\mathbf{I})$

Posterior parameter distribution by Bayes rule $p(a|b) = p(a)p(b|a)/p(b)$:

$$p(\mathbf{w}|\mathbf{x},\mathbf{y},\mathcal{M}) \;=\; \frac{p(\mathbf{w}|\mathcal{M})p(\mathbf{y}|\mathbf{x},\mathbf{w},\mathcal{M})}{p(\mathbf{y}|\mathbf{x},\mathcal{M})} \;=\; \mathcal{N}(\mathbf{w};\ \boldsymbol{\mu},\ \mathbf{\Sigma})$$

$$\mathbf{\Sigma} \;=\; \big(\sigma_{\text{noise}}^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \sigma_{\mathbf{w}}^{-2}\,\mathbf{I}\big)^{-1} \quad \text{and} \quad \boldsymbol{\mu} \;=\; \Big(\mathbf{\Phi}^\top\mathbf{\Phi} + \frac{\sigma_{\text{noise}}^2}{\sigma_{\mathbf{w}}^2}\,\mathbf{I}\Big)^{-1}\mathbf{\Phi}^\top\mathbf{y}$$

The predictive distribution is given by:

$$p(y_*|x_*,\mathbf{x},\mathbf{y},\mathcal{M}) \;=\; \mathcal{N}(y_*;\ \boldsymbol{\phi}(x_*)^\top\boldsymbol{\mu},\ \boldsymbol{\phi}(x_*)^\top\mathbf{\Sigma}\boldsymbol{\phi}(x_*) + \sigma_{\text{noise}}^2)$$