Some remarks about Bayesian Infinite Regression and Gaussian Processes

Carl Edward Rasmussen
Max Planck Institute for Biological Cybernetics
Spemann Straße 38, 72076 Tübingen, Germany
carl@tuebingen.mpg.de

Non-parametric Workshop, NIPS*03, Dec 13th
Linear in the parameters regression

Goal:
Given a data set \( \{x_c, y_c | c = 1, \ldots, n \} \), we want to predict the output \( y_* \) given a new input \( x_* \).

Model:
Consider a linear model with fixed basis functions \( \phi_1(x), \ldots, \phi_m(x) \):

\[
y = \Phi w + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_w^2 I),
\]

where all the \( \phi \)'s for each case form rows in \( \Phi \).

Prior:
Independent Gaussian prior on the weights:

\[
w \sim \mathcal{N}(0, \sigma_w^2 I).
\]
Making predictions

Under the prior, the function values are jointly Gaussian:

\[ y \sim N(\mu, Q), \quad \mu = \langle \Phi w + \epsilon \rangle = 0, \]
\[ Q = \langle yy^\top \rangle = \langle \Phi w w^\top \Phi^\top + \epsilon \epsilon^\top \rangle = \sigma_w^2 \Phi \Phi^\top + \sigma_n^2 I. \]  

(3)

Augmenting with the test case:

\[ y_{\text{aug}} = (y_1, \ldots, y_n, y_\star)^\top, \quad y_{\text{aug}} \sim N(0, Q_{\text{aug}}). \]  

(4)

Conditioning on the observed targets:

\[ y_\star|y_1, \ldots, y_n \sim N(\mu, \sigma^2), \quad \mu = Q_{\star,1:n} Q^{-1} y \]
\[ \sigma^2 = Q_{\star,\star} - Q_{\star,1:n} Q^{-1} Q_{1:n,\star}. \]  

(5)
A common choice for the basis functions $\phi_m$ is “Gaussian bumps”, centered on the training data points:

$$\phi_j(x) = \exp\left(-\left(x - x_j\right)^2\right).$$

Noise-free 95% posterior confidence region. 5 data points, known noise magnitude $\sigma_n^2 = 0.01$ and basis functions $\phi_j(x) = \exp\left(-\left(x - x_j\right)^2\right)$. 

**Gaussian bumps on the training data**

![Graph showing Gaussian bumps on training data](image)
The infinite limit

**Key idea:** Let’s put (scaled down) bumps everywhere!

The sum over basis functions becomes an integral:

\[
Q_{ij} \propto \int_{z_{\text{min}}}^{z_{\text{max}}} \phi_z(x_i)\phi_z(x_j)dz + \delta_{ij}\sigma_n^2
\]

\[
= \int_{z_{\text{min}}}^{z_{\text{max}}} \exp(- (x_i - z)^2) \exp(- (x_j - z)^2)dz + \delta_{ij}\sigma_n^2.
\]

Extending the limits to infinity, we can solve the integral:

\[
Q_{ij} \propto \exp(- (x_i - x_j)^2/2) + \delta_{ij}\sigma_n^2.
\]

The same example in the infinite case

Computational note
The computation still “only” requires inversion of an $n \times n$ matrix.
Conclusions

- Even if your model manages to explain the training points, it may still be underfitting.

- Uncertainties may be underestimated in too simple (finite, parametric) models.

- Trivial to implement for Gaussian process regression models with Gaussian noise.

- Important distinction: degenerate vs non-degenerate Gaussian process models.