GAUSSIAN PROCESSES

In these notes we will apply Gaussian random processes and Bayesian estimation theory to regression.

The result will be a method for nonlinear regression that may be viewed as a probabilistic version of _____

_____ _____ .

Bayesian Estimation

In a parameter estimation problem, we observe

$$Z \sim p(z; \theta)$$

and the objective is to estimate θ that best explains the observation.

In Bayesian estimation, we assume the parameter θ is _____ and hypothesis a _____ distribution

$$\theta \sim p(\theta).$$

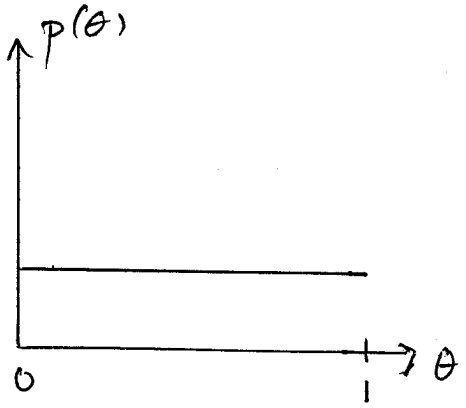
By Bayes rule,

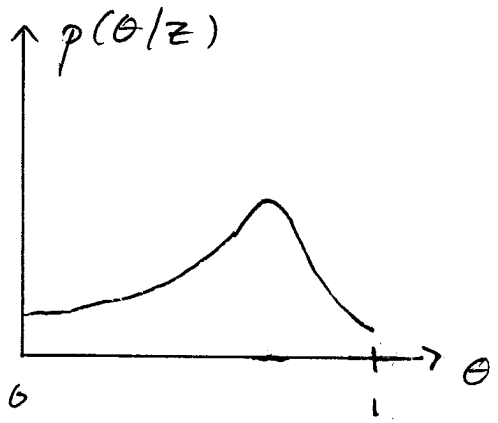
$$p(\theta \mid z) =$$

which is called the _____ distribution of θ.

All Bayesian parameter estimates are based on $p(\theta \mid z)$, e.g.,

- posterior mode: $\hat{\theta}(z) =$
- posterior mean: $\hat{\theta}(z) =$





Gaussian Processes

In most applications of Bayesian estimation, the parameter is finite dimension.

In our application, the "parameter" is the unknown regression function f

$$y = f(x) + \varepsilon$$

where

$x \in \mathbb{R}^d$

$y \in \mathbb{R}$

$\varepsilon \in \mathbb{R}$ (noise)

Therefore, a prior for f is a _____ _____, that is,

a family of random variables indexed by a potentially uncountably infinite variable, in our case

Definition | A Gaussian process is a collection of variables, any finite number of which have a multivariate Gaussian distribution.

Since Gaussian random variables are completely specified by their 1st and 2nd order statistics, a Gaussian process is completely specified by its

- mean function

$$m(x) =$$

- covariance function

$$k(x, x') =$$

Notation |

$$f(x) \sim$$

How would you generate a plot of a random function?

Note | $m(x)$ can be arbitrary, but $k(x, x')$ must satisfy

- $k(x, x') =$
- for any n, and any $x_1, \ldots, x_n \in \mathbb{R}^d$, the matrix

is _____ _____ _____.

Gaussian conditioning property

If $A \in \mathbb{R}^p$, $B \in \mathbb{R}^q$, and

$$\begin{bmatrix} A \\ B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_A \\ m_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

then

$$A \mid B = b \sim \mathcal{N}\left(m_A + \Sigma_{AB} \Sigma_{BB}^{-1} (b - m_B), \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA} \right)$$

Noise-free Case

Given

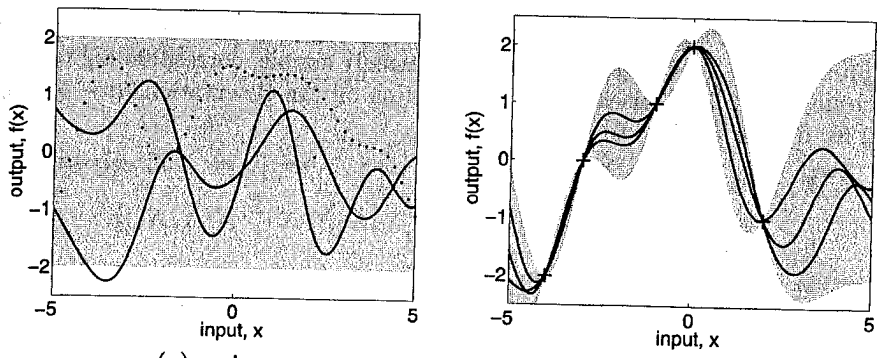
- training inputs
- training outputs
- test inputs
- mean function
- covariance function

Goal

Predict test outputs $\tilde{f} =$

Approach

- $\begin{bmatrix} \tilde{f} \\ f \end{bmatrix}$ is multivariate Gaussian distributed
- compute posterior distribution of $\tilde{f} \mid f$ using Gaussian conditioning lemma



(a), prior (b), posterior

Figure 2.2: Panel (a) shows three functions drawn at random from a GP prior; the dots indicate values of y actually generated; the two other functions have (less correctly) been drawn as lines by joining a large number of evaluated points. Panel (b) shows three random functions drawn from the posterior, i.e. the prior conditioned on the five noise free observations indicated. In both plots the shaded area represents the pointwise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region), for the prior and posterior respectively.

Rasmussen & Williams, 2006

↑ In this example, $m(x) = 0$, and the smoothness is determined by

Example

$$k(x, x') = \sigma_g^2 \exp\left\{ - \frac{\|x - x'\|^2}{2\sigma_s^2} \right\}$$

- large σ_s^2 • smooth f
- small σ_s^2 • wiggly f

Noisy case

Given

- training inputs $x_1, \ldots, x_m$
- noisy training outputs $y_i = f(x_i) + \varepsilon_i, \quad 1 \le i \le m$
- test inputs $\tilde{x}_1, \ldots, \tilde{x}_n$
- $m(x), \; k(x, x'), \; \sigma_\varepsilon^2$

Assume

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad \text{independent}$$

Goal

Predict $\tilde{f}_i = f(\tilde{x}_i), \quad 1 \le i \le n$

Approach

- $\begin{bmatrix} \tilde{f} \\ y \end{bmatrix}$ is multivariate Gaussian
- compute posterior distribution of $\tilde{f} \mid y$ using Gaussian conditioning lemma

Notation

$$\underline{X} = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}, \quad \underline{\tilde{X}} = \begin{bmatrix} \tilde{x}_1 & \cdots & \tilde{x}_n \end{bmatrix}$$

$(d \times m)$ $\qquad$ $(d \times n)$

$$f = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}, \quad \tilde{f} = \begin{bmatrix} \tilde{f}_1 \\ \vdots \\ \tilde{f}_n \end{bmatrix} = \begin{bmatrix} f(\tilde{x}_1) \\ \vdots \\ f(\tilde{x}_n) \end{bmatrix}$$

$(m \times 1)$ $\qquad$ $(n \times 1)$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad m = \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \quad \tilde{m} = \begin{bmatrix} m(\tilde{x}_1) \\ \vdots \\ m(\tilde{x}_n) \end{bmatrix}$$

$(m \times 1)$ $\qquad$ $(m \times 1)$ $\qquad$ $(n \times 1)$

$$K(\underline{X}, \underline{X}) = \left[k(x_i, x_j) \right]_{i,j=1}^{m} \qquad (m \times m)$$

$$K(\underline{X}, \underline{\tilde{X}}) = \left[k(x_i, \tilde{x}_j) \right]_{i=1, j=1}^{m, n} \qquad (m \times n)$$

$$K(\underline{\tilde{X}}, \underline{\tilde{X}}) = \left[k(\tilde{x}_i, \tilde{x}_j) \right]_{i,j=1}^{n} \qquad (n \times n)$$

Then

$$\begin{bmatrix} \tilde{f} \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \quad \end{bmatrix}, \begin{bmatrix} \quad \end{bmatrix} \right)$$

By the Gaussian conditioning lemma,

$$\tilde{f} \mid y \sim \mathcal{N}(\mu, \Sigma)$$

where

$$\mu =$$

$$\Sigma =$$

What estimates for $\tilde{f}$ does this suggest?

$$\tilde{f} \mid y \sim \mathcal{N}\Big(\tilde{m} + K(\underline{X}, \tilde{\underline{X}})^T \left[K(\underline{X}, \underline{X}) + \sigma_\varepsilon^2 I \right]^{-1} (y - m),$$
$$K(\tilde{\underline{X}}, \tilde{\underline{X}}) - K(\underline{X}, \tilde{\underline{X}})^T \left[K(\underline{X}, \underline{X}) + \sigma_\varepsilon^2 I \right]^{-1} K(\underline{X}, \tilde{\underline{X}}) \Big)$$

Since $\tilde{f} \mid y$ is symmetric, the natural estimate is

$$\hat{\tilde{f}} = \tilde{m} + K(\underline{X}, \tilde{\underline{X}})^T \left[K(\underline{X}, \underline{X}) + \sigma_\varepsilon^2 I \right]^{-1} (y - m)$$

This predictor is _____ in y

but _____ in x_i, $\tilde{x}_j$.

Does this predictor look familiar?

Connection to Kernel Ridge Regression

Assume from now on that $m(x) \equiv 0$

$(\Rightarrow \quad m, \tilde{m} = 0)$

Denote

$$\alpha = \left[K(\underline{X}, \underline{X}) + \sigma_\varepsilon^2 I \right]^{-1} y$$

Consider a single test point x $(n=1)$.

The GP prediction of $f(x)$ is

$$\hat{f}(x) =$$

This is the same as the kernel ridge regression estimate with $\lambda = \sigma_\varepsilon^2$.

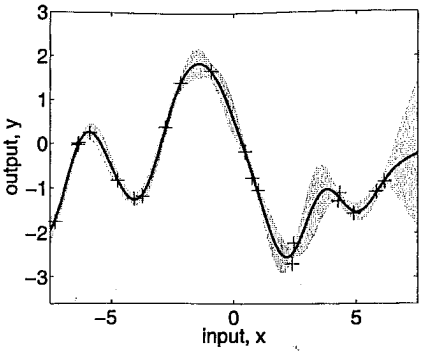
So what have we gained?

The GP predictor comes paired with a variance

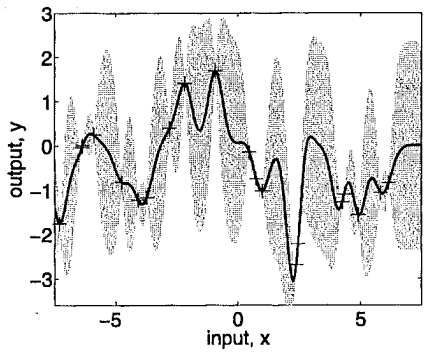
$$\text{var}(\hat{f}(x)) = k(x,x) - K(\underline{X},x)^T \left[K(\underline{X},\underline{X}) + \sigma_\varepsilon^2 I\right]^{-1} K(\underline{X},x)$$

This allows us to place a "confidence band" around the estimate whose width is, say, four times the standard deviation.

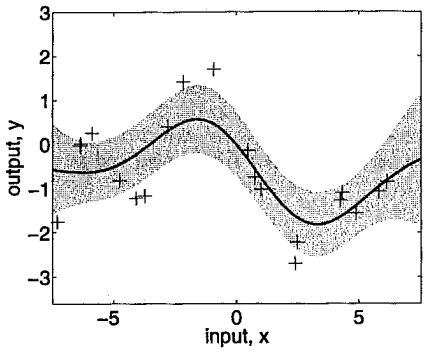
Rasmussen & Williams, 2006



(a)



(b)



(c)

	σ_ε^2	σ_s^2	σ_g^2
(a)	0.1	1	1
(b)	0.00005	0.3	1.08
(c)	0.89	3	1.16

Kernel Bayesian Linear Regression

The exact GP inference procedure, including the exact posterior $f(x) \mid y$, can be derived from a different perspective that further illuminates the connection to kernel ridge regression.

Bayesian Linear Regression

Consider a linear regression model

$$y_i = w^T x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Let's consider a Bayesian estimate of w

Data model

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \text{independent}$$

Prior

$$w \sim N(0, \sigma_w^2 I)$$

Bayes rule:

$$p(w \mid y) = \frac{p(y \mid w)\, p(w)}{p(y)}$$

- $p(w) \propto \exp\left\{ -\frac{\|w\|^2}{2\sigma_w^2} \right\}$

- $p(y \mid w) = \prod_{i=1}^{n} p(y_i \mid w)$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_\varepsilon} \exp\left\{ -\frac{(y_i - w^T x_i)^2}{2\sigma_\varepsilon^2} \right\}$$

$$\propto \exp\left\{ -\frac{\|y - \underline{X}^T w\|^2}{2\sigma_\varepsilon^2} \right\}$$

- $p(y) = \text{constant}$

$$\Rightarrow p(w \mid y) \propto \exp\left\{ -\frac{\|y - \underline{X}^T w\|^2}{2\sigma_\varepsilon^2} \right\} \exp\left\{ -\frac{\|w\|^2}{2\sigma_w^2} \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right\}$$

complete the square

where

$$\mu = \left(\underline{X}\,\underline{X}^T + \frac{\sigma_\varepsilon^2}{\sigma_w^2} I \right)^{-1} \underline{X}\, y$$

$$\Sigma = \left(\frac{1}{\sigma_\varepsilon^2}\, \underline{X}\,\underline{X}^T + \frac{1}{\sigma_w^2} I \right)^{-1}$$

Now consider a test point x. The predicted function value is

$$f(x) = w^T x$$

$$= x^T w$$

$$\Rightarrow f(x) \mid y \sim N\left(x^T \left(\underline{X}\underline{X}^T + \frac{\sigma_\varepsilon^2}{\sigma_w^2} I \right)^{-1} \underline{X} y, \right.$$

$$\left. x^T \left(\frac{1}{\sigma_\varepsilon^2} \underline{X}\underline{X}^T + \frac{1}{\sigma_w^2} I \right)^{-1} x \right)$$

matrix identities

$$\sim N\left(\sigma_w^2 x^T \underline{X} \left(\underline{X}^T \underline{X} + \frac{\sigma_\varepsilon^2}{\sigma_w^2} I \right)^{-1} y, \right.$$

$$\left. \sigma_w^2 \left[x^T x - x^T \underline{X} \left(\underline{X}^T \underline{X} + \frac{\sigma_\varepsilon^2}{\sigma_w^2} I \right)^{-1} \underline{X}^T x \right] \right)$$

- Posterior mean of $f(x) \mid y$

  $\Rightarrow$ ridge regression with $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_w^2}$

- Kernel trick

  $\Rightarrow$ GP method