Tutorial Gaussian Processes

suggested Reading on Gaussian Processes:

Bishop, section 6.4
OR
Murphy chapter 15
OR
Barber chapter 19
OR
Rasmussen, chapter 1 and 2

THE source is Rasmussen. This is STRONGLY recommended and also covers the tutorial completely. Murphy explicitly states that his chapter is a short excerpt of Rasmussen's book. Bishop is solid as always and recommended for those that stuck to Bishop anyway throughout the semester. The notation may differ a bit from the tutorial notes. Murphy is compact and easy to understand.

1 Simple Regression

We have a data set $\boldsymbol{X} \in \mathbb{R}^1$. You are given Gaussian processes $f \sim \mathcal{GP}(m, K)$ with mean function $m(x) = 0$, covariance function $K(x, x')$, and a noisy observation $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$.

Problem 1: Assume we use $K(x, x') = (xx' + 1)^2$ as our covariance function and furthermore have observations $x_1 = -\frac{1}{2}, x_2 = 2$. Write down the distribution of $p(f(x_1), f(x_2))$. What is the relationship of $f(x_1)$ and $f(x_2)$?

1 Simple Regression

We have a data set $\boldsymbol{X} \in \mathbb{R}^1$. You are given Gaussian processes $f \sim \mathcal{GP}(m, K)$ with mean function $m(x) = 0$, covariance function $K(x, x')$, and a noisy observation $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$.

Problem 1: Assume we use $K(x, x') = (xx' + 1)^2$ as our covariance function and furthermore have observations $x_1 = -\frac{1}{2}, x_2 = 2$. Write down the distribution of $p(f(x_1), f(x_2))$. What is the relationship of $f(x_1)$ and $f(x_2)$?

We have $K(x_1, x_1) = (\frac{1}{4} + 1)^2 = \frac{25}{16}$, $K(x_1, x_2) = -\frac{1}{2} \cdot 2 + 1 = 0$ and $K(x_2, x_2) = (2 \cdot 2 + 1)^2 = 25$

$$p(f(x_1), f(x_2)) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \frac{25}{16} + \sigma_y^2 & 0 \\ 0 & 25 + \sigma_y^2 \end{pmatrix}\right)$$

The covariance function is diagonal, so they are independent.

Problem 2: Now lets assume we have values $y_1 = 4, y_2 = -1$ and unknown value f_* for $x_* = 1$. We also set $\sigma_y^2 = 1$, What is the conditional distribution $\boldsymbol{f}_* | \boldsymbol{y}, \boldsymbol{X}, x_*$

Problem 2: Now lets assume we have values $y_1 = 4, y_2 = -1$ and unknown value f_* for $x_* = 1$. We also set $\sigma_y^2 = 1$, What is the conditional distribution $\boldsymbol{f}_*|\boldsymbol{y}, \boldsymbol{X}, x_*$

$$\begin{aligned}\boldsymbol{f}_*|\boldsymbol{y}, \boldsymbol{X}, x_* \sim \mathcal{N}(&\boldsymbol{\mu}_* + \boldsymbol{K}_*^T[\boldsymbol{K} + \sigma_y^2\boldsymbol{I}]^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),\\ &\boldsymbol{K}_{**} - \boldsymbol{K}_*^T[\boldsymbol{K} + \sigma_y^2\boldsymbol{I}]^{-1}\boldsymbol{K}_*).\end{aligned}$$

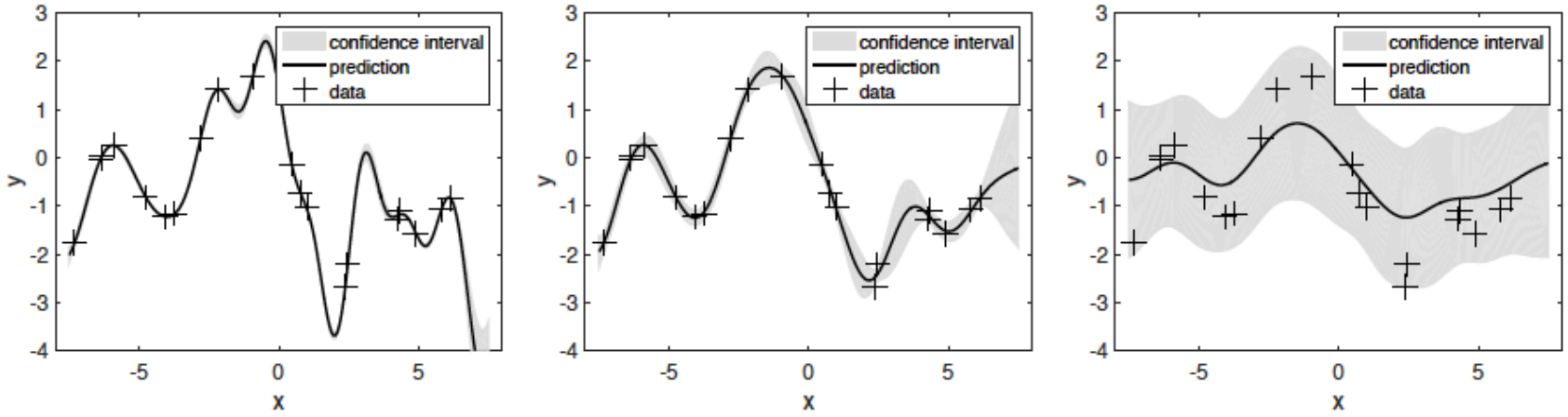
We have

$$\boldsymbol{K} + \boldsymbol{I}\sigma_y^2 = \begin{pmatrix} \frac{31}{16} & 0 \\ 0 & 26 \end{pmatrix} \quad \boldsymbol{K}_{**} = K(x_*, x_*) = (1 \cdot 1 + 1)^2 = 4 \quad \boldsymbol{K}_* = \begin{pmatrix} \frac{1}{4} \\ 9 \end{pmatrix}$$

and therefore

$$\begin{aligned}\boldsymbol{f}_*|\boldsymbol{y}, \boldsymbol{X}, x_* &\sim \mathcal{N}(\ (\tfrac{1}{4}, 9) \begin{pmatrix} \frac{16}{31} & 0 \\ 0 & \frac{1}{26} \end{pmatrix} \begin{pmatrix} 4 \\ -1 \end{pmatrix},\\ &\qquad 4 - (\tfrac{1}{4}, 9) \begin{pmatrix} \frac{16}{31} & 0 \\ 0 & \frac{1}{26} \end{pmatrix} \begin{pmatrix} 4 \\ -1 \end{pmatrix})\\ &\sim \mathcal{N}(1/4 \cdot 64/31 - 9/26, 4 - 1/4 \cdot 64/31 - 9/26)\\ &\sim \mathcal{N}(0.16997\ldots, 3.1377\ldots)\end{aligned}$$

Problem 3: We have a squared exponential kernel. With different values of σ_y^2, the GP models are shown in the figures below. Which model is best? What causes the other two to be not good? Explain your answer.



The model in the middle is best. On the left, the noise variance σ_y^2 is too small, in the right model, it is too large.

2 Weight Space and Function Space GPs are Equivalent

(Using the notation of Rasmussen (Gaussian Processes for Machine Learning (Rasmussen, Williams), (free download at www.gaussianprocess.org)).

Assume we have a dataset $\{(x_i, y_i)\}_{I=1}^n$ with $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}^1$, and a feature map $\phi : \mathbb{R}^D \to \mathbb{R}^N$. We can then form a $D \times n$ pattern matrix X and a $N \times n$ design matrix ϕ.

From Bayesian Linear Regression we know that the full Bayesian approach predictive distribution for the function value f_* of some new x_* is given by

$$f_*|x_*, X, y \sim \mathcal{N}(\sigma_n^{-2}\phi(x_*)^T A^{-1}\phi y, \phi(x_*)^T A^{-1}\phi(x_*))$$

with

$$A = \sigma_n^{-2}\phi\phi^T + \Sigma_p^{-1}$$

where σ_n is from $y = f + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_n)$ and Σ_p is from the w-prior $w \sim \mathcal{N}(0, \Sigma_p)$.

Problem 4: Kernelize this expression by rearranging it so that any dependence on X or x_* is in terms of $\phi(\ldots)^T\Sigma_p^{-1}\phi(\ldots)$ and replacing $\phi(x)^T\Sigma_p^{-1}\phi(x') = \phi(x)^T\Sigma_p^{-T/2}\Sigma_p^{-1/2}\phi(x') = \psi(x)^T\psi(x') = k(x, x')$! Show this is equivalent to Gaussian Processes with noise!

2 Weight Space and Function Space GPs are Equivalent

(Using the notation of Rasmussen (Gaussian Processes for Machine Learning (Rasmussen, Williams), (free download at www.gaussianprocess.org)).

Assume we have a dataset $\{(x_i, y_i)\}_{I=1}^n$ with $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}^1$, and a feature map $\phi : \mathbb{R}^D \to \mathbb{R}^N$. We can then form a $D \times n$ pattern matrix X and a $N \times n$ design matrix ϕ.

From Bayesian Linear Regression we know that the full Bayesian approach predictive distribution for the function value f_* of some new x_* is given by

$$f_*|x_*, X, y \sim \mathcal{N}(\sigma_n^{-2}\phi(x_*)^T A^{-1}\phi y, \phi(x_*)^T A^{-1}\phi(x_*))$$

with

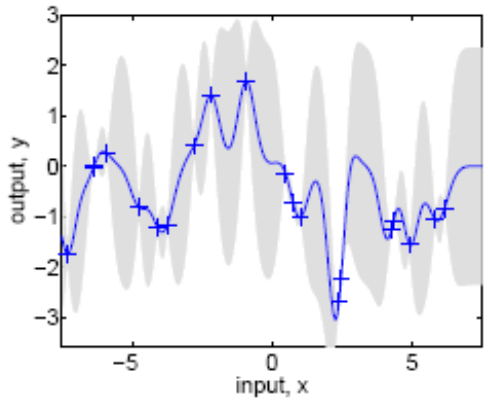
$$A = \sigma_n^{-2}\phi\phi^T + \Sigma_p^{-1}$$

where σ_n is from $y = f + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_n)$ and Σ_p is from the w-prior $w \sim \mathcal{N}(0, \Sigma_p)$.

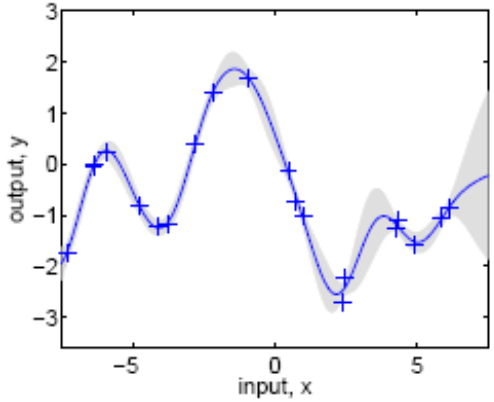
Problem 4: Kernelize this expression by rearranging it so that any dependence on X or x_* is in terms of $\phi(\ldots)^T\Sigma_p^{-1}\phi(\ldots)$ and replacing $\phi(x)^T\Sigma_p^{-1}\phi(x') = \phi(x)^T\Sigma_p^{-T/2}\Sigma_p^{-1/2}\phi(x') = \psi(x)^T\psi(x') = k(x, x')$! Show this is equivalent to Gaussian Processes with noise!

The solution *is* sections 2.1 and 2.2 of the Rasmussen book (Gaussian Processes for Machine Learning (Rasmussen, Williams) → `www.gaussianprocess.org`)

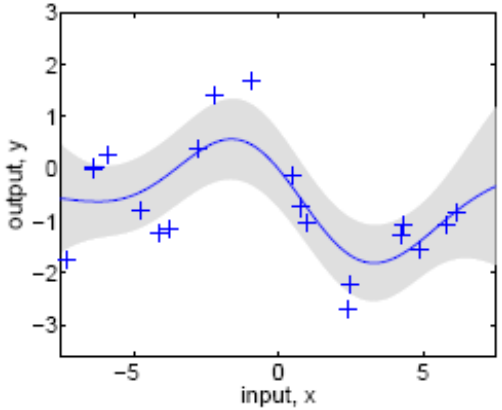
Varying the Hyperparameters



$l = 0.3$



$l = 1$



$l = 3$

The squared-exponential covariance function in one dimension:

$$k_y(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2l^2}(x_p - x_q)^2) + \sigma_n^2 \delta_{pq}$$

These kernel parameters are interpretable and can be learned from data:

l:length-scale

σ_f^2:signal variance

σ_n^2:noise variance