Gaussian Processes

Lokesh and Indra

Oct 16, 2022

GP Equations

- Training Data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$

Assumptions

- $y = f(x) + \epsilon; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$
- Prior $[f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)] \sim \mathcal{N}(0, K)$
- Then the target variable $[y_1, \cdots y_N | \mathbf{f}] \sim \mathcal{N}(\mathbf{f}, \sigma^2)$

Problem Statement

- For a test point $\mathbf{x}^*$, find the label y^*. Because noise is anywas zero-mean, we are interested to find $\Pr(f^* | \mathbf{x}^*, D)$

Write the formula for $\Pr[\mathbf{f}, f^*]$

$\Pr(f^*|\mathbf{x}^*, D)$

$$\begin{aligned}
\Pr(f^*|\mathbf{x}^*, D) &= \int_{f_1,\cdots,f_N} \Pr(f^*, f_1, \cdots, f_N|\mathbf{x}^*, D) d\mathbf{f} \\
&= \int_{\mathbf{f}} \Pr(f^*|\mathbf{f}, \mathbf{x}^*, D) \Pr(\mathbf{f}|\mathbf{x}^*, D) d\mathbf{f}
\end{aligned}$$

Data Likelihood $\Pr(\mathbf{f}|\mathbf{x}^*, D)$

Question: Can you remove some terms from the conditioning set?

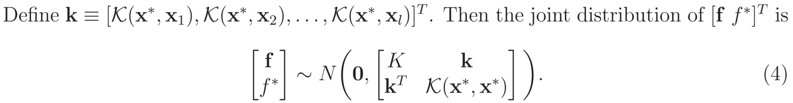
- $\Pr(\mathbf{f}|D) \propto \Pr(D|\mathbf{f})\Pr(\mathbf{f})$
- $\Pr(D|\mathbf{f}) = \Pr(\{(x_i, y_i)\}|\mathbf{f})$

Question: Simplify this further and obtain an expression for the Data likelihood. **Note** That expression should not involve $\mathbf{x}$.

Finally, by analytically evaluating the above integral, we get the data likelihood as:
$\Pr(\mathbf{f}|D) = \mathcal{N}(K(K+\sigma^2 I)^{-1}\mathbf{y}, \sigma^2 K(K+\sigma^2 I)^{-1})$

Question: When the Data Likelihood the maximum? i.e., when will the GP *perfectly* fit the dataset?

Posterior $\Pr(f^*|\mathbf{f}, \mathbf{x}^*, D)$



Define $\mathbf{k} \equiv [\mathcal{K}(\mathbf{x}^*, \mathbf{x}_1), \mathcal{K}(\mathbf{x}^*, \mathbf{x}_2), \ldots, \mathcal{K}(\mathbf{x}^*, \mathbf{x}_l)]^T$. Then the joint distribution of $[\mathbf{f}\ f^*]^T$ is

$$\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K & \mathbf{k} \\ \mathbf{k}^T & \mathcal{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right). \tag{4}$$

Some properties of Multi-variate GPs

$$P_{X,Y} = \begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right)$$

$$X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$$
$$Y \sim \mathcal{N}(\mu_Y, \Sigma_{YY})$$

$$X|Y \sim \mathcal{N}(\,\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y),\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\,)$$
$$Y|X \sim \mathcal{N}(\,\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X),\ \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\,)$$

What is the marginal $\Pr(f^*)$
What is the conditional $\Pr(f^*|\mathbf{f})$

Tying all together

$\Pr(f^*|\mathbf{x}^*, D) = \int_{\mathbf{f}} \Pr(f^*|\mathbf{f}, \mathbf{x}^*, D) \Pr(\mathbf{f}|\mathbf{x}^*, D) d\mathbf{f}$ Question: Why is the result of above integration Gaussian?

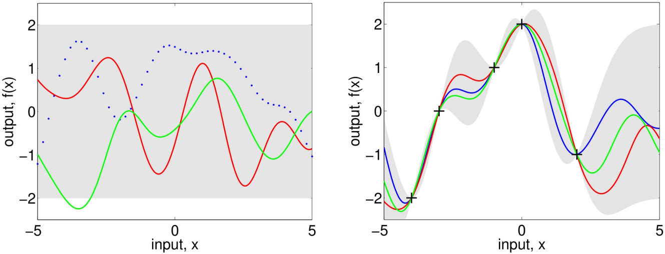
Some complicated math gives us:

$$N\left(\mathbf{k}^T\left(K+\sigma^2 I\right)^{-1}\mathbf{y},\mathcal{K}(\mathbf{x}^*,\mathbf{x}^*)-\mathbf{k}^T\left(K+\sigma^2 I\right)^{-1}\mathbf{k}\right)$$

Question: How should **k** look like for the above equation to act as 1-Nearest neighbor regressor?

Question: Do the labels **y** affect the variance of f^*?

Explain the Sampling Procedure



(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.

Give the Pseudocode used to obtain these plots.

Consider a Gaussian Process $f(x) \sim GP(0, K)$ where K is the RBF kernel $K(x_1, x_2) = exp(-\frac{1}{2}(x_1 - x_2)^2)$, where $x \in \mathbb{R}$ and mean is 0. We have one data point $D = \{x_0 = 0, y_0 = -1\}$. Answer the following :

(a) What is $-2\mu(x) + 4\sigma^2(x)$, where $\mu(x), \sigma^2(x)$ are mean and variance of the posterior distribution of $f(x)|D$ when

- $x = \sqrt{\log(4)}$
- $x \to \infty$

(b) Let $y_1 = f(x = \sqrt{log(4)})$ and $y_2 = f(x = -\sqrt{log(4)})$ be 2 Random Variables. What is $y_1 + y_2$ where $y_1, y_2 = \arg\max \Pr(y_1, y_2|D)$?

Suppose we want to sample a point $\mathbf{x}$ such that a GP that is fit on D exhibits the least variance there. Formulate the objective that gives us this.

We are estimating a 1-D regression function $f(x)$ as a Gaussian Process $GP(m(x), K(x, x'))$ where the kernel function $K(x, x') = \sigma_f^2 exp(\frac{(x-x')^2}{2\tau})$. $m(x) = 0$. Each $y_i = f(x_i) + \epsilon_i; \epsilon_i \sin N(0, \sigma^2)$. Assume Training Data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Express briefly in qualitative terms the shape of the mean of $f(x|D)$ posterior to seeing the training data D in terms of the following properties:

- Noise σ^2 is very large.
- Length scale τ changes from 10^{-4} to 10^4.