Notes on Gaussian Processes

Daniel Jiwoong Im
School of Engineering
University of Guelph
imj@uoguelph.ca

Abstract

This note was made after studying the Gaussian process (GP) using various resources. This is written as a simple summary of the GP and should not to be used as a study guide. The code for Gaussian process regression and classification are included.

1 Gaussian Distributions

Before diving into the Gaussian process, here are some of the properties about the Gaussian distribution, which will be useful for understanding the Gaussian Process.

The Gaussian distribution is a continuous probability distribution that has a bell shape probability density function. The Gaussian distribution is described by the second order statistics, the mean and the covariance,

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma)^{\frac{1}{2}}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where μ is the mean and σ^2 is the variance. For multivariate Guassian distributions,

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(\frac{-1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

where $\mathbf{x}$ is the D dimensional data vector, μ is the mean vector, and Σ is the covariance matrix. The moment and second moment of the Gaussian distribution is

$$E[\mathbf{x}] = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\int \exp\left(\frac{-1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right)(\mathbf{z}+\mu)d\mathbf{z} = \mu \tag{1}$$

$$E[\mathbf{x}\mathbf{x}^T] = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\int \exp\left(\frac{-1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right)(\mathbf{z}+\mu)^T(\mathbf{z}+\mu)d\mathbf{z} = \mu^T\mu + \Sigma \tag{2}$$

where $\mathbf{z} = \mathbf{x} + \mu$.

One useful property of the Gaussian distribution is that the product of two Gaussian distributions is Gaussian. Given two Gaussian random variable $x_1(t)$ and $x_2(t)$ with the mean and varaince corresponding to μ_1, μ_2, σ_1^2 and σ_2^2, the product of two random variables is $x_1(t)x_2(t) \sim \mathcal{N}(\mu,\sigma^2)$ such that

$$\mu = \frac{\frac{\mu_1}{2\sigma_1^2} + \frac{\mu_2}{2\sigma_2^2}}{\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{3}$$

$$\sigma_2^2 = \frac{1}{\frac{1}{\sigma^1} + \frac{1}{\sigma^2}} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{4}$$

Another very useful property is *consistency*, also known as the marginalization property. If the two samples came from the Gaussian distribution such that $(x_1, x_2) \sim N(\mu, \Sigma)$, then $x_1 \sim N(\mu_1, \Sigma_{11})$ where Σ_{11} is the relevant submatrix of Σ. In other words, if two sets of variables are jointly Gaussian, the conditional distribution of one set conditioned on the other is Gaussian. Moreover, the mariginal distribution of either set is also Gaussian as well. This gives us very handy tools for us to manipulate and write the equations in terms of block matrices. Let us re-write the data $\mathbf{x}$, mean μ, and covariance matrix Σ to be

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_a} \\ \mathbf{x_b} \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

where Λ is called the precision matrix.

Then we can re-write the mahalanobis distance of the Gaussian distribution as

$$\begin{aligned}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) = (\mathbf{x}_a-\mu_a)^T\Lambda_{aa}(\mathbf{x}_a-\mu_a) + (\mathbf{x}_a-\mu_a)^T\Lambda_{ab}(\mathbf{x}_b-\mu_b) \\ + (\mathbf{x}_b-\mu_b)^T\Lambda_{ba}(\mathbf{x}_a-\mu_a) + (\mathbf{x}_b-\mu_b)^T\Lambda_{bb}(\mathbf{x}_b-\mu_b).\end{aligned}$$

Using these partitioned form of matrices, we can define the marginal distribution and conditional distribution as $p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}|\mu_a, \Sigma_{aa})$ and $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\mu_{a|b}, \Sigma_{a|b})$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \mu_b) \tag{5}$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \tag{6}$$

2 The Gaussian Process

According to [1], the Guassian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. The Gaussian process is entirely specified by its mean and convariance as a function of the input.

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{7}$$

$$k(\mathbf{x}, \mathbf{x}^{'}) = \mathbb{E}[f(\mathbf{x} - m(\mathbf{x}))(f(\mathbf{x}^{'}) - m(\mathbf{x}^{'}))] \tag{8}$$

The prior of the Gaussian process is a distribution over the function space. In practice, we do not much have the prior knowledge about the mean of $f(\mathbf{x})$, so we set the mean to be zero, $\mathbb{E}[f(\mathbf{x})] = 0$.

The Gaussian process can be viewed as a kernel method and it can be related to different machine learning techniques such as the radial basis function, ridge regression, and Bayesian linear regression. As many textbook introduced the Gaussian process starting from the Bayesian linear regression, we can see the relationship between the two. Suppose $\mathbf{y}(\mathbf{x}) = \Phi(\mathbf{x})^T W$ and $W \sim \mathcal{N}(0, \Sigma_p)$ where Φ is the function that maps the D-dimensional input vector to M-dimensional feature space. Then, the second order statistics are

$$\mathbb{E}[\mathbf{y}(\mathbf{x})] = \Phi\mathbb{E}(\mathbf{y}(\mathbf{x})) = 0 \tag{9}$$

$$\mathbb{E}[\mathbf{y}(\mathbf{x})\mathbf{y}(\mathbf{x}^T)] = \Phi\mathbb{E}[WW^T]\Phi^T = \Phi\Sigma_p\Phi^T \tag{10}$$

From this derivation, we can see that $\mathbf{y}(\mathbf{x})$ is Gaussian that has a zero mean and covariance $\Phi\Sigma_p\Phi^T$. Hence, this is a particular example of the Gaussian process.

The reason why the Gaussian process is a kernel method is because we specify the covariance function $k(\mathbf{x}, \mathbf{x}^{'})$ to be a kernel function. For example, we can set covariance function to be

$$\text{Linear Kernel: } k(\mathbf{x}, \mathbf{x}^{'}) = \mathbf{x}\mathbf{x}^{'} \tag{11}$$

$$\text{Linear Exponential: } k(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(-\frac{1}{2}|\mathbf{x} - \mathbf{x}^{'}|\right) \tag{12}$$

$$\text{Squared Exponential: } k(\mathbf{x}, \mathbf{x}^{'}) = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}^{'}\|^2\right). \tag{13}$$

The choice of the kernel function depends on the user and the user decide the smoothness of the function.

We can now easily sample a function with the zero mean and a covariance function $k(\mathbf{x}, \mathbf{x}^{'})$

$$\mathbf{f}^* \sim \mathcal{N}(\mathbf{0}, k(\mathbf{x}^*, \mathbf{x}^*))$$

where $\mathbf{x}^*$ are samples. One thing about the Gaussian processes is that the formulation itself is not that hard, but the concept of function-space view could be little bit overwhelming at first. This is because even though we say that we have a prior distribution over the function space, we only considered functions at the points that we have in our dataset, but I just embraced it.

3 Gaussian Process Regression

Until now, we only talked about drawing a sample function based on our prior knowledge. Here, we incorporate the dataset to predict the function and unseened points $\mathbf{x}^*$. Suppose we have the noise-free dataset such that $\{(\mathbf{x}_i, f_i)|\forall i = 1 \cdots n)\}$ and we like to sample functions at points $\{\mathbf{x}_j^*|\forall j = 1 \cdots m\}$. In order to sample, we have to define a joint distribution based on the discussion from the above.

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(X,X) & k(X,\mathbf{x}^*) \\ k(\mathbf{x}^*,X) & k(\mathbf{x}^*,\mathbf{x}^*) \end{bmatrix}\right)$$

We can predict f^* at $\mathbf{x}^*$ conditioning on our observation $(X, \mathbf{f})$, we will use the conditional Gaussian probability from Equation 5. Then we get,

$$\mathbf{f}^*|\mathbf{x}^*, X, \mathbf{f}^* \sim \mathcal{N}\left(k(\mathbf{x}^*, X)k(X,X)^{-1}\mathbf{f}, k(\mathbf{x}^*,\mathbf{x}^*) - k(\mathbf{x}^*,X)k(X,X)^{-1}k(X,\mathbf{x}^*)\right).$$

Now we will repeat the same process but assume that there is noise in our dataset and we will also assume that the noise is indepedent of the data distribution. Let's define the dataset to be $\{(\mathbf{x}_i, y_i)|\forall i = 1 \cdots n)\}$. We use y instead of f because it is not a true function anymore due to the noise, i.e $y = f(\mathbf{x}) + \epsilon, \epsilon\, \mathcal{N}(0, \delta)$. Then, the joint distribution of $(\mathbf{y}, \mathbf{f}^*)$ is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(X,X) + \delta^2 I & k(X,\mathbf{x}^*) \\ k(\mathbf{x}^*,X) & k(\mathbf{x}^*,\mathbf{x}^*) \end{bmatrix}\right)$$

Similarly, the conditional distribution becomes

$$\mathbf{f}^*|\mathbf{x}^*, X, \mathbf{y}^* \sim \mathcal{N}\left(k(\mathbf{x}^*, X)\left[k(X,X)^{-1} + \delta^2 I\right]^{-1}\mathbf{y}, k(\mathbf{x}^*,\mathbf{x}^*) - k(\mathbf{x}^*,X)\left[k(X,X) + \delta^2 I\right]^{-1}k(X,\mathbf{x}^*)\right).$$

One of the reasons why we say that the radial basis kernel method is a special case of the Gaussian process is because we can reformulate the mean of f^* to be

$$\bar{f}^* = k(\mathbf{x}^*, X)\left[k(X,X)^{-1} + \delta^2 I\right]^{-1}\mathbf{y} = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}^*)$$

where $\alpha_i = \left[k(X,X)^{-1} + \delta^2 I\right]^{-1}\mathbf{y}$. Now if our kernel function is a squared exponential function, then it becomes a radial basis kernel formulae

$$\bar{f}^* = \sum_{i=1}^{n} \alpha_i \exp\left(\frac{-1}{2}\|\mathbf{x}_i - \mathbf{x}^*\|^2\right).$$

Lastly, we will talk about marginal likelihood $p(\mathbf{y}|\mathbf{x})$, which is an integral of the likelihood and prior,

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f}$$

The prior of $\mathbf{f}|X \sim \mathcal{N}(\mathbf{0}, k(X,X))$ is Gaussian and the likelihood $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \delta^2 I)$ is Gaussian so $p(\mathbf{y}|X)$ is Gaussian as shown in Equation 3. Then, log of $p(\mathbf{y}|X)$ is

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T\left(k(X,X) + \delta^2 I\right)\mathbf{y} - \frac{1}{2}\log|k(X,X) + \delta^2 I| - \frac{n}{2}\log 2\pi.$$

4 Gaussian Process Classification

When we do classification with the Gaussian process, where the predictions take the form of discrete values (class labels), we cannot use the likelihood function as the Gaussian distribution as for the regression problem. This means that we cannot make use of the conjugate-prior to the classification. Thus, both the GP regression and classification are function approximators but this makes the Gaussian process classification to be more challenging than the regression.

Suppose that $(\mathbf{x}_i, y_i)$ are the data and label for all $i = 1, \cdots, n$ and $(\mathbf{x}^*)$ is the data point that we want to predict. Then, the conditional probability of f^* becomes

$$p(f^*|X, \mathbf{y}, \mathbf{x}^*) = \int p(f^*|X, \mathbf{x}^*, \mathbf{f}) p(\mathbf{f}|X, \mathbf{y}) d\mathbf{f} \tag{14}$$

which is very similar to the regression except that we incorporate the labels $\mathbf{y}$. Plus, the probability prediction based on f^* becomes

$$p(y^* = 1|X, \mathbf{y}, \mathbf{x}^*) = \int p(y = 1|\mathbf{x}^*, f^*)) p(f^*|X, \mathbf{y}, \mathbf{x}^*) df^* \tag{15}$$

As mentioned, $p(y = 1|f^*(\mathbf{x}))$ is typically some regression function within a range between $[0, 1]$. Also note that $p(\mathbf{f}|X, \mathbf{y})$ and $p(y = 1|f^*(\mathbf{x}))$ are non-gaussian functions.

The marginal probability $p(\mathbf{y}|x)$ is expressed as

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|X) d\mathbf{f} = \int \exp\left(\Phi(\mathbf{f})\right) d\mathbf{f}$$

In order to compute the probability for each class $p(y|\mathbf{f})$ we choose the probability function in the range of $[0, 1]$. In practice, we use

$$\text{Linear regression: } p(\mathbf{y}|\mathbf{f}) = \frac{1}{1 + \exp\left(-\mathbf{f}\right))}$$

$$\text{Probit regression: } p(\mathbf{y}|\mathbf{f}) = \int_{-\infty}^{f} \mathcal{N}(x|0, 1) dx$$

The difficulties in the Gaussian process classification comes from Equation 14 and Equation 15, because we cannot analytically compute the integral. Two ways to approximate the integral: by using the Laplace approximation and the expectation propagation.

4.1 Laplace Approximation

Since $p(\mathbf{f}|X, \mathbf{y})$ is non-Gaussian, the Laplace approximation method tries to approximate $p(\mathbf{f}|X, \mathbf{y})$ using the Gaussian function $q(\mathbf{f}|X, \mathbf{y})$. This can be done by approximating the second order Taylor expansion of $\log p(\mathbf{f}|X, \mathbf{y})$ around the maximum of the posterior $\hat{\mathbf{f}}$

$$q(\mathbf{f}|X, \mathbf{y}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, A^{-1})$$

where $A = -\nabla\nabla \log p(\mathbf{f}|X, \mathbf{y})$ at $\hat{\mathbf{f}}$. The maximum of the posterior of $\hat{\mathbf{f}}$ can be found using Newton's iteration method. Details can be found in [2]. Similarly, we approximate the marginal probability $p(\mathbf{y}|X)$ by approximating with Gaussian distribution. Then the marginal probability in Equation 4 becomes:

$$p(\mathbf{y}|X) \simeq q(\mathbf{y}|X) = \exp\left(\Phi(\hat{\mathbf{f}})\right) \int \exp\left(-\frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^T A (\mathbf{f} - \hat{\mathbf{f}}) d\mathbf{f}\right)$$

Then, the approximated log likeliehood of marginal probability becomes

$$\log q(\mathbf{y}|X) = (-\frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^T A (\mathbf{f} - \hat{\mathbf{f}})) + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2} \log |B|$$

where $B = I + W^{\frac{1}{2}} K W^{\frac{1}{2}}$, $W = -\nabla\nabla p(\mathbf{y}|\mathbf{f})$ and K is covariance matrix for $p(\mathbf{f}|X)$.

References

[1] Christopher M. Bishop, Pattern Recognition and Machine learning, Springer-Verlag New York.

[2] Carl Edward and Christopher K. I. Williams, Gaussian Proecsses for Machine learning, The MIT Press, CAmbridge, Massachusetts.