Cogsci 118A: Natural Computation I Lecture 16

Gaussian Processes and Kernel Functions

Lecturer: Angela Yu *Scribe: Joseph Schilz*

Lecture Summary

1. *Exploring the Gaussian Process*
   We explore the properties of Gaussian Processes through the Gaussian Process applet.
2. *Gaussian Processes Continued*
   We continue our formal exploration of Gaussian Processes. We include a proof that linear basis function regression is a special case of the Gaussian Process and explore methods for validating kernel functions.

1 Exploring the Gaussian Process

The following considerations arise in exploring the Gaussian Process applet.

Consider Figure 1. How much should our observation at x_1 influence our predicted mean at x_2 and x_3? The kernel function answers this question by transforming the Euclidean distance between observations and predictions into some new distance metric. Figure 1 demonstrates a Gaussian kernel. The influence of our observation on our prediction varies with the height of the graph ???.

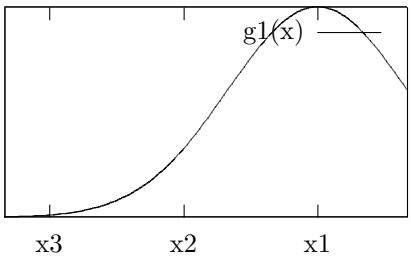

Figure 1: We observe x_1. The kernel function determines how much our predictions at x_2 and x_3 should be influenced by our observation.

Analyzing the form of the kernel, we find that p_3 represents the independent noise or the observation noise. We see that p_2 parameterizes the magnitude of influence of observations upon predictions. This parameter determines the width of the non-normalized Gaussian pictured in Figure 1. We see that p_1 increases the variance between data points.

Other kernel functions use transform Euclidean distance into other metrics. We may also explore periodic and polynomial kernel functions via the applet. Note that under the polynomial kernel function influence varies with the product of our observed x, rather than with distance between points ???.

2 Gaussian Processes Continued

Recall from previous lectures that there is a formal equivalence between linear basis function regression and Gaussian Processes. As we will show, linear basis function regression is a special case of a Gaussian Processes.

Recall also that Gaussian Processes are defined by the joint probability of our observations $\mathbf{y}$. That joint probability is Gaussian, with mean $\mathbf{0}$. That is:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{0}, K),$$

where y may be written implictly or explictly as a function of $\mathbf{x}$, as in $\mathbf{y} = [y_1(\mathbf{x}_1), y_2(\mathbf{x}_2), \ldots, y_n(\mathbf{x}_n)]^T$. The covariance matrix K is of size $n \times n$. In Gaussian Processes, the covariance matrix might also be called the Gramm matrix, a smoother matrix, or simply the covariance. It is of the form:

$$K = \begin{bmatrix} k_{11} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{n1} & \cdots & k_{nn} \end{bmatrix} \text{ where } k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Recall that a covariance matrix must be symmetric and positive semidefinite. That is, $\Sigma_{ij} = \Sigma_{ji}$ and for all properly sized $\mathbf{z}$ we have $\mathbf{z}^T \Sigma \mathbf{z} \geq 0$. Then K must be symmetric and positive semidefinite. So we definite our *kernel function* $k(\mathbf{x}, \mathbf{x}')$, such that for any finite collection of inputs $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $k(\mathbf{x}, \mathbf{x}')$ induces a cov matrix that is symmetric and positive semidefinite.

Recall from previous lectures the following theorem:

Theorem 1 *A Gramm matrix is symmetric and positive semidefinite if and only if there exists a feature map $\boldsymbol{\phi}(\mathbf{x})$ such that $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}')$. Where $\boldsymbol{\phi}(\mathbf{x})$ can be infinite dimensional.*

This means that a Gramm matrix may be equivalently represented by either a kernel function or by a basis function. As a concrete example, we now explore the kernel function induced by linear regression basis functions.

2.1 Linear Basis Function Regression as a Gaussian Process

Recall the following formulation of linear basis function regression:

$$y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \text{ with prior } p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbb{I}).$$

Equivalently we may write write:

$$\mathbf{y} = \Phi\mathbf{w}, \text{ where the design matrix } \Phi = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

We note that $p(\mathbf{y}|\mathbf{w})$ is Gaussian because it's a linear function of a Gaussian random varianble $\mathbf{w}$. It has mean and covariance:

$$\mathbb{E}[y] = \mathbb{E}[\Phi\mathbf{w}] = \Phi\mathbb{E}[\mathbf{w}] = \mathbf{0} \text{ by } \mathbb{E}[\mathbf{w}] = \mathbf{0}.$$

$$\text{cov}[y] = \mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] - \mathbb{E}[\mathbf{y}]^2 = \mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbb{E}\left[\Phi\mathbf{w}\mathbf{w}^T\Phi^T\right] = \Phi\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\right]\Phi^T = \Phi[\alpha^{-1}\mathbb{I}]\Phi^T = \alpha^{-1}\Phi\Phi^T.$$

Then $p(\mathbf{y}) = \mathcal{N}(\mathbf{0}, K)$ where $K = \alpha^{-1}\Phi\Phi^T$.

Note that k_{ij} is the ith row of Φ times the jth column of Φ^T, times the scalar α^{-1}. Then $k_{ij} = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$. Note also that α is positive. Then we may define:

$$\boldsymbol{\phi}'(\mathbf{x}) \triangleq \frac{1}{\sqrt{\alpha}}\boldsymbol{\phi}(\mathbf{x}) \text{ which yields } k_{ij} = \boldsymbol{\phi}'(\mathbf{x}_i)^T\boldsymbol{\phi}'(\mathbf{x}_j).$$

We have shown that the joint probability of $\mathbf{y}$ under linear basis function regression is Gaussian with mean $\mathbf{0}$ and covariance K, and we constructed a ϕ such that $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$. Then by the definition of a Gaussian Process and Theorem 1 we find that linear basis function regression is a case of the Gaussian Process.

2.2 Validating Kernel Functions

There are two methods to show that a candidate kernel function k is valid. We may use Theorem 1, as we did above, by identifying a vector of basis functions $\boldsymbol{\phi}$ such that $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$. We refer to this as method (A) below. Alternately, we may verify directly that K is symmetric and positive semidefinite. We refer to this approach as method (B).

We use these two methods to demonstrate valid extensions of kernel functions. In demonstrating the following properties, we take k_1, k_2 as known valid kernel functions. These methods of extension are useful in constructing new kernels and in validating new kernel functions.

1. Where $k(\mathbf{x}, \mathbf{x}') = c \cdot k_1(\mathbf{x}, \mathbf{x}')$ for $c \geq 0$, k is a valid kernel function.

   Proof By Theorem 1 we know $k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ for some $\boldsymbol{\phi}(\mathbf{x})$. Then define $\boldsymbol{\phi}' = \sqrt{c}\boldsymbol{\phi}(\mathbf{x})$. Then $k = \boldsymbol{\phi}'(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ is a valid kernel function by Theorem 1. □

2. $k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x})$ where $f(\mathbf{x})$ is any function.

   Proof By Theorem 1 we know $k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ for some $\boldsymbol{\phi}(\mathbf{x})$. Then define $\boldsymbol{\phi}' = f(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})$???. Then $k = \boldsymbol{\phi}'(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ is a valid kernel function by Theorem 1. □

3. $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$

   Proof We proceed by showing that the Gramm matrix K induced by k is symmetric and positive semidefinite.

   Showing that K is symmetric follows trivially from the symmetry of K_1 and K_2. Noting that $K = K_1 + K_2$, for an arbitrary $\mathbf{z}$ we find that $\mathbf{z}^T K\mathbf{z} = \mathbf{z}^T(K_1 + K_2)\mathbf{z} = \mathbf{z}^T K_1\mathbf{z} + \mathbf{z}^T K_2\mathbf{z}$. And by $\mathbf{z}^T K_1\mathbf{z} \geq 0$, $\mathbf{z}^T K_2\mathbf{z} \geq 0$ we have $\mathbf{z}^T K\mathbf{z} \geq 0$. □

We will show in our homework that the following kernel functions are valid:

1. $k = k_1 \cdot k_2$
2. $k = q(k_1)$ where q is a polynomial with nonnegative coefficients

3. $k = \exp(k_1)$
4. $k = k_1(\phi(\mathbf{x}), \phi(\mathbf{x}'))$
5. $k = \mathbf{x}^T A \mathbf{x}'$
6. $k = k_a(x_a, x'_a) + k_b(x_b, x'_b)$
7. $k = k_a(x_a, x'_a) \cdot k_b(x_b, x'_b)$

2.3 Choosing Parameters for the Gaussian Process

We may require need to select one or more parameters in the construction of our Gramm matrix, depending upon our choice of kernel function k. One such commonly used kernel function is given by:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2\sigma^2}\right].$$

How do we choose our parameters σ? In general, let $\boldsymbol{\theta}$ be our vector of parameters. In this case $\boldsymbol{\theta} = [\sigma]$. Then we may perform iterative gradient descent as follows:

$$\hat{\theta}_i^{t+1} = \hat{\theta}_i^t - \frac{d}{d\theta_i}\left(-\log p(\mathbf{t}, \boldsymbol{\theta})\right)???.$$

We shall not pursue this algorithm further in class, homework, or test. However, you may find it useful in the future.

2.4 Gaussian Processes Prediction Summary

Unlike linear basis function regression which has a fixed number of basis functions, Gaussian Processes grow in complexity as more data is acquired. The lack of basis functions makes Gaussian Processes an example of a nonparametric Bayesian model.