$\mathcal{GP}$ Write Up

September 26, 2018

Convolutions on the Kernel matrix

1. Introduction

Predictive mean for a GP is:

$$\mu_L = \mathbf{K}_*(\mathbf{K_X} + \sigma_\mathbf{n}^\mathbf{2}\mathbf{I})^{-1}\mathbf{y} \tag{1}$$

the parameters for the covariance function $k(.,.)$ are optimized by maximizing the log marginal likelihood given by:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_\mathbf{n}^\mathbf{2}I)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{2}$$

Equation (1) involves computing inverses and this is computationally expensive and we need to circumvent this problem subject to some assumptions on the training data.

We will see one of the ways of avoiding the inverse when the training data is on a grid and the kernel is stationary.

Say:

$$\mathbf{v} = (\mathbf{K_X} + \sigma_\mathbf{n}^\mathbf{2}\mathbf{I}^{-1})\mathbf{y} \tag{3}$$

We can avoid calculating the inverse by solving the linear system:

$$(\mathbf{K_X} + \sigma_\mathbf{n}^\mathbf{2}\mathbf{I})\mathbf{v} = \mathbf{y} \tag{4}$$

for $\mathbf{v}$.

2. Convolutions on a grid

When the training data $X \in \mathbb{R}^{n\times n}$, say $n = 2$, and with a stationary kernel

$$k(\mathbf{x_i}, \mathbf{x_j}) = f'(\mathbf{x_i} - \mathbf{x_j}) = f(x_i^{(0)} - x_j^{(0)}, x_i^{(1)} - x_j^{(1)}) \tag{5}$$

equation (4) can be written as:

$$\mathbf{y} = \begin{bmatrix} f(0,0) & f(0,-1) & f(-1,0) & f(-1,-1) \\ f(0,1) & f(0,0) & f(-1,1) & f(-1,0) \\ f(1,0) & f(1,-1) & f(0,0) & f(0,-1) \\ f(1,1) & f(1,0) & f(0,1) & f(0,0) \end{bmatrix} \begin{bmatrix} v_{0,0} \\ v_{0,1} \\ v_{1,0} \\ v_{1,1} \end{bmatrix} + \sigma_n^2\mathbf{I}.\,\mathbf{v} \tag{6}$$

This matrix multiplication can be represented as a convolution:

$$\mathbf{y} = \mathbf{K}' * \mathbf{v}' + \sigma_n^2\mathbf{I}.\,\mathbf{v} \tag{7}$$

where:

$$\mathbf{K}' = \begin{bmatrix} f(-1,-1) & f(-1,0) & f(-1,1) \\ f(0,-1) & f(0,0) & f(0,1) \\ f(1,-1) & f(1,0) & f(1,1) \end{bmatrix} \tag{8}$$

and

$$\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{1,0} \\ v_{0,1} & v_{0,0} \end{bmatrix} \tag{9}$$

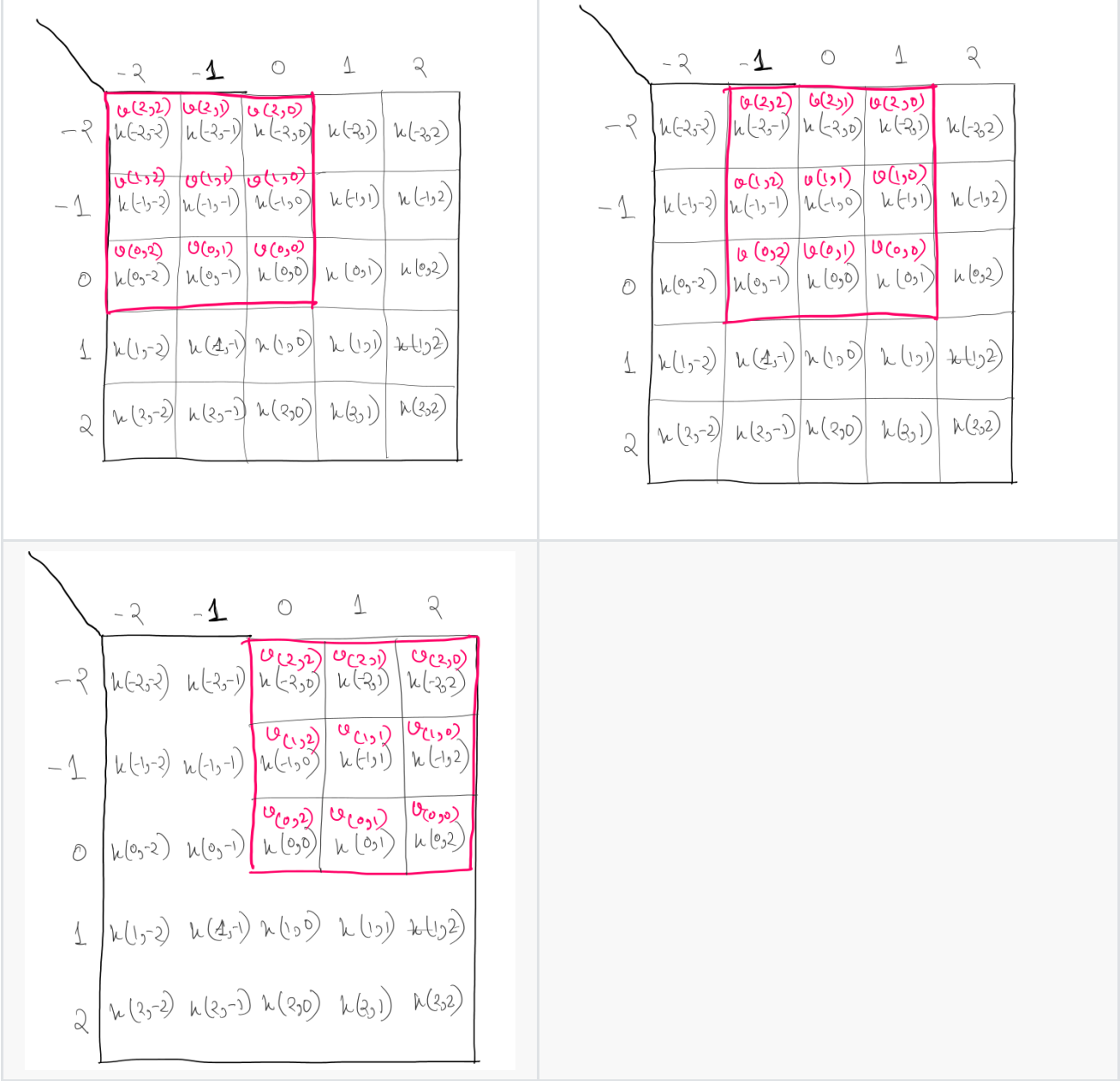
Lets see convolutions on a grid where $n = 3$:

like in equation (8)

$$\mathbf{K}' = \begin{bmatrix} f(-2,-2) & f(-2,-1) & f(-2,0) & f(-2,1) & f(-2,2) \\ f(-1,-2) & f(-1,-1) & f(-1,0) & f(-1,1) & f(-1,2) \\ f(0,-2) & f(0,-1) & f(0,0) & f(0,1) & f(0,2) \\ f(1,-2) & f(1,-1) & f(1,0) & f(1,1) & f(1,2) \\ f(2,-2) & f(2,-1) & f(2,0) & f(2,1) & f(2,2) \end{bmatrix} \tag{10}$$

and like in equation (9)

$$\mathbf{v}' = \begin{bmatrix} v_{2,2} & v_{2,1} & v_{2,0} \\ v_{1,2} & v_{1,1} & v_{1,0} \\ v_{0,2} & v_{0,1} & v_{0,0} \end{bmatrix} \tag{11}$$



Therefore

$$\mathbf{y} = \mathbf{K}' * \mathbf{v}' \tag{12}$$

where $\mathbf{y} \in \mathbb{R}^{n \times n}$ and $\mathbf{K}' \in \mathbb{R}^{(2n-1)\times(2n-1)}$. The computational complexity is still $O(n^4)$ however we have decreased the space complexity from $O(n^4)$ to $O(n^2)$.

3. Effect of Missing observations

Consider the scenario where some observations are missing on the grid and we fill the missing observations with samples from $\mathcal{N}(0, \sigma_w^2 I)$ where $\sigma_w \to \infty$. Lets say we have M observations and W missing observations for a total of N observations.

$$\mathbf{K}_N = \begin{bmatrix} \mathbf{K}_M & \mathbf{K}_{MW} \\ \mathbf{K}_{MW}^T & \mathbf{K}_W \end{bmatrix} + \begin{bmatrix} \sigma_m^2 I_M & \mathbf{0} \\ \mathbf{0} & \sigma_w^2 I_W \end{bmatrix} \tag{13}$$

from equation (3):

$$(\mathbf{K}_N + \mathbf{D}_N)^{-1}\mathbf{y} = \mathbf{v} \tag{14}$$

we can use the block matrix inversion algorithm to see what this inverse leads to:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(I - D^{-1}CA^{-1}B)^{-1}D^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (I - D^{-1}CA^{-1}B)^{-1}D^{-1} \end{bmatrix} \tag{15}$$

where:

$$A = K_M + \sigma_M^2 I_M \tag{16}$$

$$B = K_{MW} \tag{17}$$

$$C = K_{MW}^T \tag{18}$$

$$D = K_W + \sigma_W^2 I_W = \sigma_W^2(\sigma_W^{-2} K_W + I_W) \tag{19}$$

$$\lim_{\sigma_W \to \infty} D^{-1} = 0 \tag{20}$$

Therefore equation (3) becomes

$$(\mathbf{K}_M + \sigma_M^2 I_M)^{-1}\mathbf{y}_M = \mathbf{v}_M \tag{21}$$

and

$$\mathbf{y}_M = (\mathbf{K}_M + \sigma_M^2 I_M)\mathbf{v}_M \tag{22}$$

from equation (1) the predictive mean for the entire grid will be:

$$\mu_L = \mathbf{K}_{NM}\mathbf{v}_M \tag{23}$$

4. Loss Function

For a function:

$$f(x) = \frac{1}{2}x^T Ax - x^T b + c \tag{24}$$

if A is symmetric the stationary point is at:

$$Ax = b \tag{25}$$

and if A is positive definite then the solution to equation (26) is the minimum. So the solution to equation (26) can be obtained by minimizing equation (24).

So the optimization problem is:

$$min_v \frac{1}{2}\mathbf{v}^T\mathbf{K}\mathbf{v} - \mathbf{v}^T\mathbf{y} \tag{26}$$

which simplifies to:

$$\frac{1}{2}\mathbf{y}'_m(\mathbf{K}_M + \mathbf{D}_M)^{-1}\mathbf{y}'_m - \mathbf{y}'_m(\mathbf{K}_M + \mathbf{D}_M)^{-1}\mathbf{y}_m \tag{27}$$