Sparse Gaussian

January 5, 2017

1 mathematical notations

We first give brief description of mathematical notations will be used through out the project.

The original data set will be denoted as $\mathcal{D}$ which consists of N d-dimensional vectors $\boldsymbol{X} = \{\boldsymbol{x}^{(i)} = (x_1, \dots, x_d) \,|\, i = 1, \dots, N\}$. Let the new input data be $\boldsymbol{x}^* = (x_1^*, \dots, x_d^*)$. The pseudo input data set is denoted as $\bar{\mathcal{D}}$ consists of $\bar{\boldsymbol{X}} = \{\bar{\boldsymbol{x}}^{(i)} = (x_1, \dots, x_d) \,|\, i = 1, \dots, M\}$. $\boldsymbol{X}$ is paired with target $\boldsymbol{Y} = (y^{(1)}, \dots, y^{(N)})$, notice that $y^{(i)}$ are scalars. $\boldsymbol{x}^*$ is paired with new target y^*. The underlining latent function is denoted as $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$ and the pseudo one is $\bar{\boldsymbol{f}}$. A Gaussian distribution is denoted as $\mathcal{N}(\boldsymbol{f}|\boldsymbol{m}, \boldsymbol{V})$ with mean $\boldsymbol{m}$ and variance $\boldsymbol{V}$.

2 sparse Gaussian process

We first give a zero mean Gaussian prior over the underlining latent function: $p(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_N)$ where $\boldsymbol{K}_N$ is our kernel matrix with elements given by, $[\boldsymbol{K}_N]_{ij} \equiv K_{\boldsymbol{x}^{(i)}\boldsymbol{x}^{(j)}} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$: Notice that this is the case that we have same number of $\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}$. In case of different sizes, we use $\boldsymbol{K}_{NM}$, i.e. N rows for the first input matrix, M rows for the second input matrix.

$$K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = c \exp[-\frac{1}{2} \sum_{k=1}^{D} b_k (x_k^{(i)} - x_k^{(j)})^2], \quad \boldsymbol{\theta} \equiv \{c, \boldsymbol{b}\}, \tag{1}$$

where $\boldsymbol{\theta}$ is the hyperparameters. We provide noises to $\boldsymbol{f}$ such that $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \boldsymbol{I})$. By integrating out the latent function we have the marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K}_N + \sigma^2 \boldsymbol{I}) \tag{2}$$

For prediction, the new input $\boldsymbol{x}^*$ conditioning on the observed data and hyperparameters. Let write the joint probability first

$$p(y^*, \boldsymbol{y}|\boldsymbol{x}^*, \mathcal{D}, \boldsymbol{\theta}) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{\boldsymbol{x}^*\boldsymbol{x}^*} + \sigma^2 & \boldsymbol{K}_{\boldsymbol{x}^*N} \\ \boldsymbol{K}_{\boldsymbol{x}^*N}^T & \boldsymbol{K}_N + \sigma^2 \boldsymbol{I} \end{pmatrix}\right), \tag{3}$$

where $\boldsymbol{K}_{\boldsymbol{x}^*N} = (K(\boldsymbol{x}^*, \boldsymbol{x}^{(1)}), \dots, K(\boldsymbol{x}^*, \boldsymbol{x}^{(N)}))$, i.e. $[\boldsymbol{K}_{\boldsymbol{x}^*N}]_i = K(\boldsymbol{x}^*, \boldsymbol{x}^{(i)})$, and $K_{\boldsymbol{x}^*\boldsymbol{x}^*} = K(\boldsymbol{x}^*, \boldsymbol{x}^*)$. Now we can condition on $\boldsymbol{y}$ and get

$$\begin{aligned} &p(y^*|\boldsymbol{y}, \boldsymbol{x}^*, \mathcal{D}, \boldsymbol{\theta}) \\ &= \mathcal{N}(y^*|\boldsymbol{K}_{\boldsymbol{x}^*N}(\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}^T, K_{\boldsymbol{x}^*\boldsymbol{x}^*} + \sigma^2 - \boldsymbol{K}_{\boldsymbol{x}^*N}(\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{K}_{\boldsymbol{x}^*N}^T). \end{aligned} \tag{4}$$

For detailed proof, check Theorem 4.3.1 in Murphy's machine learning a probabilistic perspective.

Now we consider pseudo input $\bar{\boldsymbol{X}}$. Everything still holds except that there are no noises in it. The new input and target pair $(\boldsymbol{x}^*, y^*)$ is replaced by one of the actually data set and targets pairs $(\boldsymbol{x}^{(i)}, y_i)$. We therefore just use $\bar{\boldsymbol{f}}$ represents the pseudo outputs and $\bar{\boldsymbol{\theta}}$, and the single point likelihood is given by

$$p(y|\boldsymbol{x}, \bar{\boldsymbol{f}}, \bar{\boldsymbol{X}}) = \mathcal{N}(y|\boldsymbol{K}_{\boldsymbol{x}M}\boldsymbol{K}_M^{-1}\bar{\boldsymbol{f}}, K_{\boldsymbol{x}\boldsymbol{x}} + \sigma^2 - \boldsymbol{K}_{\boldsymbol{x}M}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{\boldsymbol{x}M}^T), \tag{5}$$

where $\boldsymbol{K}_{\boldsymbol{x}M} = (K(\boldsymbol{x}, \bar{\boldsymbol{x}}^{(1)}), \ldots, K(\boldsymbol{x}, \bar{\boldsymbol{x}}^{(M)}))$, i.e. $[\boldsymbol{K}_{\boldsymbol{x}M}]_i = K(\boldsymbol{x}, \bar{\boldsymbol{x}}^{(i)})$. As the target data are i.i.d given the inputs, the complete data likelihood is given by

$$p(\boldsymbol{y}|\boldsymbol{X}, \bar{\boldsymbol{f}}, \bar{\boldsymbol{X}}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x}^{(i)}, \bar{\boldsymbol{f}}, \bar{\boldsymbol{X}}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{K}_{NM}\boldsymbol{K}_M^{-1}\bar{\boldsymbol{f}}, \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I}), \tag{6}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda}), \lambda_i = K_{\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)}} - \boldsymbol{K}_{\boldsymbol{x}^{(i)}M}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{\boldsymbol{x}^{(i)}M}^T$, is a $N \times N$ diagonal matrix, and $[\boldsymbol{K}_{NM}]_{ij} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Together with a Gaussian prior, $p(\bar{\boldsymbol{f}}|\bar{\boldsymbol{X}}) = \mathcal{N}(\bar{\boldsymbol{f}}|\boldsymbol{0}, \boldsymbol{K}_M)$, integrate over Eq.6 we have the SPGP marginal likelihood over pseudo inputs

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}, \bar{\boldsymbol{X}}) &= \int p(\boldsymbol{y}|\boldsymbol{X}, \bar{\boldsymbol{f}}, \bar{\boldsymbol{X}}) p(\bar{\boldsymbol{f}}|\bar{\boldsymbol{X}}) \, \mathrm{d}\bar{\boldsymbol{f}} \\ &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K}_{NM}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN} + \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I}). \end{aligned} \tag{7}$$

Same as we have done from Eq.3 to Eq.4, we first write the joint probability of $y^*, \boldsymbol{y}$

$$\begin{aligned} &p(y^*, \boldsymbol{y}|\boldsymbol{x}^*, \boldsymbol{X}, \bar{\boldsymbol{X}}) \\ &= \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{\boldsymbol{x}^*\boldsymbol{x}^*} + \sigma^2 & \boldsymbol{K}_{\boldsymbol{x}^*M}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN} \\ (\boldsymbol{K}_{\boldsymbol{x}^*M}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN})^T & \boldsymbol{K}_{NM}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN} + \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I} \end{pmatrix}\right), \end{aligned} \tag{8}$$

where $\boldsymbol{K}_{\boldsymbol{x}^*M} = (K(\boldsymbol{x}^*, \bar{\boldsymbol{x}}^{(1)}), \ldots, K(\boldsymbol{x}^*, \bar{\boldsymbol{x}}^{(M)}))$, i.e. $[\boldsymbol{K}_{\boldsymbol{x}^*M}]_i = K(\boldsymbol{x}^*, \bar{\boldsymbol{x}}^{(i)})$. From now on we let

$$\boldsymbol{Q}_{\boldsymbol{X},\boldsymbol{X}'} \equiv \boldsymbol{Q}(\boldsymbol{X}, \boldsymbol{X}') = \boldsymbol{K}_{\boldsymbol{X}M}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{M\boldsymbol{X}'} \tag{9}$$

$$\boldsymbol{Q}_N = \boldsymbol{K}_{NM}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN}, \tag{10}$$

Also, remember that here N and M represents input and pseudo input data set, matrices, as input matrices of $\boldsymbol{K}$, respectively. And after conditioning on $\boldsymbol{y}$, we have the SPGP predictive distribution

$$p(y^*|\boldsymbol{y}, \boldsymbol{x}^*, \boldsymbol{X}, \bar{\boldsymbol{X}}) = \mathcal{N}(\mu^*, \sigma^{*2}) \tag{11}$$

$$\begin{aligned} \mu^* &= \boldsymbol{Q}_{\boldsymbol{x}^*N}(\boldsymbol{Q}_N + \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y} \\ \sigma^{*2} &= K_{\boldsymbol{x}^*\boldsymbol{x}^*} - \boldsymbol{Q}_{\boldsymbol{x}^*N}(\boldsymbol{Q}_N + \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Q}_{N\boldsymbol{x}^*} + \sigma^2. \end{aligned} \tag{12}$$

The pseudo input $\bar{\boldsymbol{C}}$ and hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \sigma^2\}$, this can be done by maximizing Eq.7.

Some simplification for matrix inversion. First from matrix inversion lemma

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{U}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{A}^{-1} \tag{13}$$

$$\det(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^T) = \det(\boldsymbol{A})\det(\boldsymbol{B})\det(\boldsymbol{B}^{-1} + \boldsymbol{U}^T\boldsymbol{A}^{-1}\boldsymbol{U}), \tag{14}$$

we can rewrite following

$$
\begin{aligned}
&(\boldsymbol{K}_{NM}\boldsymbol{K}_M^{-1}\boldsymbol{K}_{MN} + \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1} && (15)\\
&= (\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1} - (\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}_{NM}\boldsymbol{B}^{-1}\boldsymbol{K}_{MN}(\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}, && (16)
\end{aligned}
$$

where $\boldsymbol{B} = \boldsymbol{K}_M + \boldsymbol{K}_{MN}(\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}_{NM}$. Now matrix inversion only happens to $(\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}$ which is $\mathcal{O}(N)$ as it is diagonal. Now Eq.12 become

$$
\begin{aligned}
\mu^* &= \boldsymbol{K}_{\boldsymbol{x}^*M}\boldsymbol{B}^{-1}\boldsymbol{K}_{MN}(\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}\\
\sigma^{*2} &= K_{\boldsymbol{x}^*\boldsymbol{x}^*} - \boldsymbol{K}_{\boldsymbol{x}^*M}(\boldsymbol{K}_M^{-1} - \boldsymbol{B}^{-1})\boldsymbol{K}_{M\boldsymbol{x}^*} + \sigma^2.
\end{aligned} \tag{17}
$$

3 implementation

Rewrite

$$
\sigma^2\boldsymbol{\Gamma} = \boldsymbol{\Lambda} + \sigma^2\boldsymbol{I}, \tag{18}
$$

and suppressing data dependency of Eq.7, we have

$$
p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{Q}_N + \sigma^2\boldsymbol{\Gamma}). \tag{19}
$$

We maximize $\mathcal{L} = -\log p(\boldsymbol{y})$

$$
\mathcal{L} = \frac{1}{2}(\log(\det(\boldsymbol{Q}_N + \sigma^2\boldsymbol{\Gamma})) + \boldsymbol{y}(\boldsymbol{Q}_N + \sigma^2\boldsymbol{\Gamma})^{-1}\boldsymbol{y}^T + N\log(2\pi)). \tag{20}
$$

Let

$$
\begin{aligned}
\mathcal{L}_1 &= \log(\det(\boldsymbol{Q}_N + \sigma^2\boldsymbol{\Gamma})) && (21)\\
\mathcal{L}_2 &= \boldsymbol{y}(\boldsymbol{Q}_N + \sigma^2\boldsymbol{\Gamma})^{-1}\boldsymbol{y}^T. && (22)
\end{aligned}
$$

Use matrix inversion lemma again, we have

$$
\begin{aligned}
\mathcal{L}_1 &= \log(\det(\boldsymbol{K}_M + \sigma^{-2}\boldsymbol{K}_{MN}\boldsymbol{\Gamma}^{-1}\boldsymbol{K}_{NM})\det(\boldsymbol{K}_M^{-1})\det(\sigma^2\boldsymbol{\Gamma}))\\
&= \log(\det(\boldsymbol{A})) - \log(\det(\boldsymbol{K}_M)) + \log(\det(\boldsymbol{\Gamma})) + (N-M)\log(\sigma^2) && (23)\\
\mathcal{L}_2 &= \sigma^{-2}\boldsymbol{y}(\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\boldsymbol{K}_{NM}\boldsymbol{A}^{-1}\boldsymbol{K}_{MN}\boldsymbol{\Gamma}^{-1})\boldsymbol{y}^T && (24)\\
&= \sigma^{-2}(||\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{y}^T||^2 - ||\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{K}_{NM})^T\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{y}^T||^2) && (25)
\end{aligned}
$$

where $\boldsymbol{A} = \sigma^2\boldsymbol{K}_M + \boldsymbol{K}_{MN}\boldsymbol{\Gamma}^{-1}\boldsymbol{K}_{NM}$. The final negative log marginal likelihood is

$$
\mathcal{L} = \frac{1}{2}(\mathcal{L}_1 + \mathcal{L}_2 + N\log(2\pi)). \tag{26}
$$

3.1 matrix derivatives

Let $\boldsymbol{A}$ be a matrix with underlining parameter θ. The derivative of the inverse matrix w.r.t θ is

$$
\frac{\partial}{\partial\theta}\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}\frac{\partial\boldsymbol{A}}{\partial\theta}\boldsymbol{A}^{-1}, \tag{27}
$$

where the partial derivative takes elementwise. If $\boldsymbol{A}$ is positive definite symmetric, the derivative of the log determinant is

$$\frac{\partial}{\partial \theta} \log(\det(\boldsymbol{A})) = \mathrm{tr}(\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial \theta}) \tag{28}$$

First ignoring the noise variance σ^2, do partial derivative on $\boldsymbol{\theta}$, we have

$$\dot{\mathcal{L}}_1 = \mathrm{tr}(\boldsymbol{A}^{-\frac{1}{2}} \dot{\boldsymbol{A}} \boldsymbol{A}^{-\frac{T}{2}}) - \mathrm{tr}(\boldsymbol{K}_M^{-\frac{1}{2}} \dot{\boldsymbol{K}}_M \boldsymbol{K}_M^{-\frac{T}{2}}) + \mathrm{tr}(\boldsymbol{\Gamma}^{-\frac{1}{2}} \dot{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}^{-\frac{1}{2}}) \tag{29}$$