Sparse Gaussian

January 2, 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 $\boldsymbol{X} = \{\boldsymbol{x}^{(i)} = (x_1, \ldots, x_d) \mid i = 1, \ldots, N\}$. Let the new input data be $\boldsymbol{x}^* = (x_1^*, \ldots, x_d^*)$. The pseudo input data set is denoted as $\bar{\mathcal{D}}$ consists of $\bar{\boldsymbol{X}} = \{\bar{\boldsymbol{x}}^{(i)} = (x_1, \ldots, x_d) \mid i = 1, \ldots, M\}$. $\boldsymbol{X}$ is paired with target $\boldsymbol{Y} = (y^{(1)}, \ldots, y^{(N)})$, 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, $[\boldsymbol{K}_N]_{nn'} = K(\boldsymbol{x}, \boldsymbol{x}')$:

$$K(\boldsymbol{x}, \boldsymbol{x}') = c\exp[-\frac{1}{2}\sum_{i=1}^{D} b_i(x_i^{(n)} - x_i^{(n')})^2], \quad \boldsymbol{\theta} \equiv \{c, \boldsymbol{b}\}, \tag{1}$$

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

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

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

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

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

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

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