Tutorial Week 11

COMP9418 – Advanced Topics in Statistical Machine Learning, 17s2, UNSW Sydney

Instructor: Edwin V. Bonilla

Last Update: Thursday 5th October, 2017 at 14:12

1. This question is concerned with binary classification problems where we are given input-output observations $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, with each $\mathbf{x} \in \mathbb{R}^D$, $y \in \{-1, +1\}$. We denote the inputs compactly with the $D \times N$ matrix $\mathbf{X}$ and the outputs with the $N \times 1$ vector $\mathbf{y}$.

   Consider a Gaussian process (GP) prior $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$, which when realised on the observed features induces a Gaussian prior over the N latent function values $\mathbf{f}$:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}), \tag{1}$$

   where $\mathbf{f}$ is the N-dimensional column vector of latent function values, each corresponding to an observed label, i.e. $\mathbf{f} = (f_1, \ldots, f_N)^T$; and $\mathbf{K}$ is the covariance matrix obtained by evaluating the covariance function at all pairwise input training points, i.e. $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X}; \boldsymbol{\theta})$.

   As usual, a suitable likelihood model for binary classification is given by the Bernoulli distribution:

$$p(y|f(\mathbf{x})) = \sigma(f(\mathbf{x}))^{\mathbb{I}(y=+1)}(1 - \sigma(f(\mathbf{x})))^{\mathbb{I}(y=-1)}, \tag{2}$$

   where $\sigma(f)$ is a sigmoid function such as the logistic sigmoid:

$$\sigma(f) = \frac{1}{1 + \exp(-f)}. \tag{3}$$

   (a) Show that, assuming iid observations, the likelihood of the GP binary classification model can be written as:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} \sigma(y^{(i)} f_i). \tag{4}$$

   (b) Explain what it means, for the model defined by Equations (1) and (4), for the posterior to be analytically intractable.

   (c) Assume that you approximate the true posterior using a Gaussian distribution $p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \approx q(\mathbf{f}|\mathbf{X}, \mathbf{y})$, where $q(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}|\mathbf{b}, \boldsymbol{\Sigma})$ is your approximate posterior. Derive an expression for the posterior predictive distribution for a new datapoint $\mathbf{x}_\star$, i.e. $p(f_\star|\mathbf{X}, \mathbf{y}, \mathbf{x}_\star)$.

   (d) Given the above approximation, explain how to compute the predictive probability $p(y_\star = +1|\mathbf{X}, \mathbf{y}, \mathbf{x}_\star)$.

2. The subset of regressors (SR) approximation relies upon the inducing-variable approach. Let us denote the M inducing variables with $\mathbf{u} = (u_1, \ldots, u_M)$. Recall that these variables

are in the same space as f, i.e. they are actual function values. Aditionally, we denote the set $\mathbf{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}\}$ as the corresponding inducing inputs.

The SR approximation assumes the following covariance function:

$$\kappa_{\mathrm{SR}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \kappa(\mathbf{x}^{(i)}, \mathbf{Z})\mathbf{K}_{\mathrm{zz}}^{-1}\kappa(\mathbf{Z}, \mathbf{x}^{(j)}), \tag{5}$$

where $\kappa(\mathbf{x}, \mathbf{Z})$ computes the kernel between $\mathbf{x}$ and the set $\mathbf{Z}$ and similarly for $\kappa(\mathbf{Z}, \mathbf{x})$ and $\mathbf{K}_{\mathrm{zz}} = \kappa(\mathbf{Z}, \mathbf{Z}; \boldsymbol{\theta})$. Under the standard GP prior and Gaussian likelihood with isotropic noise with variance σ_n^2,

(a) show that the predictive mean and covariance for the SR model are given by:

$$\mathbb{E}[\mathbf{f}_\star | \mathbf{y}, \mathbf{X}] = \mathbf{K}_{\star\mathrm{z}} \left(\mathbf{K}_{\mathrm{zx}}\mathbf{K}_{\mathrm{xz}} + \sigma_n^2 \mathbf{K}_{\mathrm{zz}}\right)^{-1} \mathbf{K}_{\mathrm{zx}}\mathbf{y}, \tag{6}$$

$$\mathbb{C}\mathrm{ov}[\mathbf{f}_\star | \mathbf{y}, \mathbf{X}] = \sigma_n^2 \mathbf{K}_{\star\mathrm{z}}(\mathbf{K}_{\mathrm{zx}}\mathbf{K}_{\mathrm{xz}} + \sigma_n^2 \mathbf{K}_{\mathrm{zz}})^{-1}\mathbf{K}_{\mathrm{z}\star}, \tag{7}$$

where $\mathbf{K}_{\star\mathrm{z}} = \kappa(\mathbf{X}_\star, \mathbf{Z}; \boldsymbol{\theta})$; $\mathbf{K}_{\mathrm{z}\star} = \mathbf{K}_{\star\mathrm{z}}^T$; $\mathbf{K}_{\mathrm{xz}} = \kappa(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$; and $\mathbf{K}_{\mathrm{zx}} = \mathbf{K}_{\mathrm{xz}}^T$.

(b) Show that the marginal likelihood can be written as:

$$\mathcal{L}_{\mathrm{SR}} = -\frac{1}{2}\Big\{ \log \left|\mathbf{K}_{\mathrm{zx}}\mathbf{K}_{\mathrm{xz}} + \sigma_n^2 \mathbf{K}_{\mathrm{zz}}\right| - \log |\mathbf{K}_{\mathrm{zz}}| + \frac{1}{\sigma_n^2}\mathbf{y}^T\mathbf{y} \tag{8}$$

$$- \frac{1}{\sigma^2}\mathbf{y}^T\mathbf{K}_{\mathrm{xz}}(\mathbf{K}_{\mathrm{zx}}\mathbf{K}_{\mathrm{xz}} + \sigma_n^2\mathbf{K}_{\mathrm{zz}})^{-1}\mathbf{K}_{\mathrm{zx}}\mathbf{y} + N\log(2\pi)\Big\}. \tag{9}$$

Comment on the time complexity of:

i. The mean and variance of the predictive distribution of the SR model.

ii. The marginal likelihood of the SR model.

iii. The gradients of the marginal likelihood wrt the inducing inputs in the SR model.