Tutorial Week 10

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

Instructor: Edwin V. Bonilla

Last Update: Thursday 28th September, 2017 at 15:25

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

1. Consider the following linear-in-the-parameters model:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Sigma}_w), \tag{1}$$

$$p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}^T\mathbf{w}, \sigma^2\mathbf{I}), \tag{2}$$

where $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{X})$ is the $D' \times N$ feature matrix, with D' being the number of (nonlinear) basis functions. The predictive distribution for the function value at a new $\mathbf{x}_\star$, i.e. $f_\star = \mathbf{w}^T\boldsymbol{\phi}_\star$, is given by:

$$p(f_\star|\mathbf{x}_\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_\star|\sigma^{-2}\boldsymbol{\phi}_\star^T\mathbf{A}^{-1}\mathbf{\Phi}\mathbf{y}, \boldsymbol{\phi}_\star^T\mathbf{A}^{-1}\boldsymbol{\phi}_\star), \tag{3}$$

where: $\boldsymbol{\phi}_\star = \boldsymbol{\phi}(\mathbf{x}_\star)$, and $\mathbf{A} = (\frac{1}{\sigma^2}\mathbf{\Phi}\mathbf{\Phi}^T + \mathbf{\Sigma}_w^{-1})$.

   (a) Show that Equation (3) can be written as:

$$p(f_\star|\mathbf{x}_\star, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_\star|\mathbf{k}_\star^T\widetilde{\mathbf{K}}^{-1}\mathbf{y}, k_{\star\star} - \mathbf{k}_\star^T\widetilde{\mathbf{K}}^{-1}\mathbf{k}_\star), \tag{4}$$

   where $\mathbf{k}_\star = \mathbf{\Phi}^T\mathbf{\Sigma}_w\boldsymbol{\phi}_\star$, $k_{\star\star} = \boldsymbol{\phi}_\star^T\mathbf{\Sigma}_w\boldsymbol{\phi}_\star$, and $\widetilde{\mathbf{K}} = \mathbf{\Phi}^T\mathbf{\Sigma}_w\mathbf{\Phi} + \sigma^2\mathbf{I}$.

   (b) Discuss the computational complexity of Equations (3) and (4).

   (c) Discuss how equation (4) allows us to have a non-linear regression model without computing feature vectors explicitly.

2. Here we want to model the data with a zero-mean Gaussian process prior with covariance $\kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the hyperparameters. Assuming iid observations corrupted by Gaussian noise with variance σ^2, this translates into the following model:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}), \tag{5}$$

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I}), \tag{6}$$

where $\mathbf{K}$ denotes the covariance matrix obtained by evaluating the covariance function at all pairwise training inputs and $\mathbf{f}$ is the $N \times 1$ vector of latent function variables.

   (a) Show that the expectation of the predictive distribution for a new test point $\mathbf{x}_\star$ is given by:

$$\mathbb{E}[f(\mathbf{x}_\star)] = \mathbf{k}_\star^T\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \tag{7}$$

   where $\mathbf{k}_\star$ is the $N \times 1$ vector obtained by evaluating the covariance function between all training points and the test point. Show that this expectation can be expressed as (i) a linear combination of the N observed outputs, and (ii) a linear combination of N kernel functions. Explain what each of these cases implies.

(b) We know that the product of two Gaussian distributions satisfies the following property:

$$\mathcal{N}(\mathbf{x}|\mathbf{m}_1, \mathbf{\Sigma}_1)\mathcal{N}(\mathbf{x}|\mathbf{m}_2, \mathbf{\Sigma}_2) = Z_c^{-1}\mathcal{N}(\mathbf{x}|\mathbf{m}_c, \mathbf{\Sigma}_c), \tag{8}$$

where $Z_c^{-1} = \mathcal{N}(\mathbf{m}_1|\mathbf{m}_2, \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)$. Show that the marginal likelihood of the model defined in Equations (5) and (6) is:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I}). \tag{9}$$

(c) Explain how the log marginal likelihood $\log p(\mathbf{y})$, where $p(\mathbf{y})$ is defined as in Equation (9), can be used for hyperparameter learning and how it naturally balances between a model-fitting term and a penalty term, which should generally avoid over-fitting.

(d) Give the time complexity and memory complexity of the mean prediction and the marginal likelihood computations in equations (7) and (9) and the implication of these when scaling up to large datasets.