Gaussian Processes for Machine Learning Chapter 2 Summary

CSCI-GA 3033 Bayesian Machine Learning
New York University
Yilun Kuang

Oct 24, 2022

2. 1 Weight-Space View

For a linear regression model $y = f(\mathbf{x}) + \epsilon = \mathbf{x}^\top \mathbf{w} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$, we have the posterior distribution $p(\mathbf{w}|X, y) \sim \mathcal{N}(\overline{\mathbf{w}} = \frac{1}{\sigma_n^2} A^{-1} X y, A^{-1})$, where $A = \sigma_n^{-2} X X^\top + \Sigma_p^{-1}$. We also have the predictive distribution $p(f(\mathbf{x}_*)|\mathbf{x}_*, X, \mathbf{y}) = \mathcal{N}(\frac{1}{\sigma_n^2} \mathbf{x}_*^\top A^{-1} X \mathbf{y}, \mathbf{x}_*^\top A^{-1} \mathbf{x}_*)$.

If we let $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$, then the predictive distribution is given by $f(\mathbf{x}_*)|\mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}(\phi_*^\top \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \mathbf{y}, \phi_*^\top \Sigma_p \phi_* - \phi_*^\top \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \Phi^\top \Sigma_p \phi_*)$.

2.2 Function-Space View

Consider the Bayesian linear regression model $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ with prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$. Then we have the mean and covariance function

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] = \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')] = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}')$$

we say $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ as a Gaussian process, completely determined by the mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$.

Choose the RBF kernel $k(\mathbf{x}_p, \mathbf{x}_q) = \exp(-\frac{1}{2}|\mathbf{x}_p - \mathbf{x}_q|^2)$. Consider the model $y = f(\mathbf{x}) + \epsilon$. Then we have the covariance function $\text{cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_n^2 \delta_{pq}$. The predictive equation for Gaussian process regression is given by $\mathbf{f}_*|X, \mathbf{y}, X_* \sim \mathcal{N}(\overline{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))$, where

$$\overline{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y}$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*$$

2.3 Varying the Hyperparameters

The squared-exponential covariance function in one dimension has the following form:

$$k_y(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x_p - x_q)^2) + \sigma_n^2 \delta_{pq}$$

The length-scale hyperparameter ℓ affects variances in the output prediction. Hyperparameters σ_f^2 and σ_n^2 can also be set by optimizing marginal likelihood.

2.4 Decision Theory for Regression

Given the predictive distribution, we would like to compute a point estimation for decision. We can choose y such that

$$y_{\text{optimal}}|\mathbf{x}_* = \text{argmin}_{y_{\text{guess}}} \int \mathcal{L}(y_*, y_{\text{guess}}) p(y_*|\mathbf{x}_*, \mathcal{D}) dy_*$$

2.5 An Example Application

See textbook.