Chapter 15
Gaussian processes

15.1 Introduction

In supervised learning, we observe some inputs $\boldsymbol{x}_i$ and some outputs y_i. We assume that $y_i = f(\boldsymbol{x}_i)$, for some unknown function f, possibly corrupted by noise. The optimal approach is to infer a *distribution over functions* given the data, $p(f|\mathcal{D})$, and then to use this to make predictions given new inputs, i.e., to compute

$$p(y|\boldsymbol{x},\mathcal{D}) = \int p(y|f,\boldsymbol{x})p(f|\mathcal{D})\mathrm{d}f \tag{15.1}$$

Up until now, we have focussed on parametric representations for the function f, so that instead of inferring $p(f|\mathcal{D})$, we infer $p(\boldsymbol{\theta}|\mathcal{D})$. In this chapter, we discuss a way to perform Bayesian inference over functions themselves.

Our approach will be based on **Gaussian processes** or **GP**s. A GP defines a prior over functions, which can be converted into a posterior over functions once we have seen some data.

It turns out that, in the regression setting, all these computations can be done in closed form, in $O(N^3)$ time. (We discuss faster approximations in Section 15.6.) In the classification setting, we must use approximations, such as the Gaussian approximation, since the posterior is no longer exactly Gaussian.

GPs can be thought of as a Bayesian alternative to the kernel methods we discussed in Chapter 14, including L1VM, RVM and SVM.

15.2 GPs for regression

Let the prior on the regression function be a GP, denoted by

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), \kappa(\boldsymbol{x},\boldsymbol{x}')) \tag{15.2}$$

where $m(\boldsymbol{x}$ is the mean function and $\kappa(\boldsymbol{x},\boldsymbol{x}')$ is the kernel or covariance function, i.e.,

$$m(\boldsymbol{x} = \mathbb{E}[f(\boldsymbol{x})] \tag{15.3}$$

$$\kappa(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}) - m(\boldsymbol{x}))^T] \tag{15.4}$$

where κ is a positive definite kernel.

15.3 GPs meet GLMs

15.4 Connection with other methods

15.5 GP latent variable model

15.6 Approximation methods for large datasets