Summary 9: Regression

Susan Cherry

The sections of this book describe both the function-space and weight-space view of regression. Section 2.1 discusses the Bayesian treatment of the linear regression with Gaussian noise. the model is as follows: $f(x) = x^T w$, $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. The paper presents the likelihood, marginal likelihood, posterior, and predictive distribution. The mean of the posterior distribution is the maximum a posteriori (MAP) estimate. While this linear model is easy to implement and interpret, it suffers from limited flexibility- if the relationship between input and outputs isn't actually linear, the model will perform poorly.

In order to allow for greater flexibility, the inputs can be projected into a higher dimensional space using basis functions. The linear model can then be applied in this higher dimensional space instead of directly on the inputs. To do this, the authors introduce $\phi(x)$ which maps a D-dimensional input vector into an N dimensional features space. Now the model is $f(x) = \phi(x)^T w$. The new predictive distribution is shown. A kernel, which is $k(x, x') = \phi(x)\phi(x')$ is introduced and the authors explain the kernel trick which means that if an algorithm is defined only in terms of inner produces in the input space, it can be lifted to the higher dimensional feature space by replacing the original inner produces with k(x,x').

Next the paper moves to talking about the Function-Space view, which is an equivalent way of viewing regression but this time by inferring inference directly in the function space. A Gaussian Process is used to describe a distribution over functions, where a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. A simple example of a Gaussian process can be obtained from the Bayesian linear regression model $f(x) = \phi(x)^T w$ with the prior $w \sim N(0, \Sigma_p)$.

We usually not interested in drawing random functions from the prior, but instead want to incorporate the knowledge that the training set gives us about the function. As a result, the next section describes prediction, first with noise free observations. The noise free predictive distribution is specified. Then it moves on to talking about prediction using noisy observations. This is important because we realistically will only have noisy versions of function values. The predictive distribution for the noisy observations is derived. The exact correspondence with the weight space view is explained. The predictive distribution for a single point shows that the mean prediction is a linear combination of observations y, which is often referred to as a linear predictor. It can also be viewed as a linear combination of n kernel functions each centered on a training point. This section concludes by providing a practical implementation of Gaussian process regression.

Section 2.3 discusses varying the hyperparameters. The length scale is set short and set long and the results are compared. The short scale length means that the error bars in Figure 2.5 grow rapidly away from the data points. In the too long length-scale model, the data is explained by a slowly varying function with a lot of noise. In Chapter 5 (which I didn't read) the authors discuss generalizing the marginal likelihood to higher dimensions, which allows us to score various models. The marginal likelihood gives a clear preferences for the (l, σ_f, σ_n) over the two alternative models.