Summary of "Gaussian Processes for Machine Learning"

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 26 February, 2014

Rasmusssen and Williams (2006) present two views of Gaussian processes. In general, a Gaussian process (GP) defines a probability distribution over functions. Gaussian process regression allows inference in this function space (the "function-space" view). The weight-space view of GPs extends the general Bayesian linear model to a high-dimensional feature space. In this feature space it is sometimes possible to use the "kernel trick" to simplify computations in high dimensional settings.

Starting with the basic Bayesian linear model with Gaussian noise

$$\begin{aligned} y &= f(x) + \epsilon \\ f(x) &= x'w \\ \epsilon &\sim N(0, \sigma^2), \end{aligned}$$

the weight space view of Gaussian process regression allows the inputs x to be projected into feature space. This offers an advantage when, for example, features for a classification problem may not be linearly separable in their original space, but could be when the projection is applied. To accomplish this, we introduce a function $\phi(x)$ to project the D-dimensional input into a N dimensional space. This mapping $\phi(x)$ replaces x in the above model so that

$$f(x) = \phi(x)'w.$$

The predictive distribution f_* becomes

$$\begin{aligned} f_*|x_*, X, y \sim\ & N(\phi(x_*)'\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}y, \\ & \phi(x_*)'\Sigma_p\phi_* - \phi(x_*)'\Sigma_p\Phi(K + \sigma^2 I)^{-1}\Phi'\Sigma_p\phi(x_*)). \end{aligned}$$

In cases where an algorithm is defined solely by inner products in input space, we can simplify the computation of f_* by defining $k(x_1, x_2) = \phi(x_1)'\Sigma_p\phi(x_2) = \Sigma_p^{1/2}\phi(x_1) \cdot \Sigma_p^{1/2}\phi(x_2)$. Then, we can replace occurrences of inner products by $k(x_1, x_2$ (the "kernel trick").

Gaussian process regression can also be understood according to the function space view, as a collection of random variables with a joint Gaussian distribution. However, we may not know the full covariance matrix of the distribution (if we did, we would have a full Gaussian distribution rather than a GP). In this case, we define a covariance function using the same kernel as above:

$$\text{cov}(f(x_1), f(x_2)) = k(x_1, x_2) = \exp(-\frac{1}{2}|x_1 - x_2|^2).$$

This covariance function defines a distribution over functions (hence, "function space"). It allows us to sample a subset of input points X_*, fill in the covariance matrix for these points, and generate a Gaussian vector

$$f_* \sim^N (0, K(X_*, X_*)).$$

We can extend this to the using of training and test sets, as above. The outcome of GP inference, then, is a posterior over the function space and a covariance between candidate functions. The GP is a powerful tool that can also be extended to regression with multiple target variables and a number of other applications.