Gaussian Process Regression

Based on Rasmussen & Williams (2006)

Chandler Lester

What is a Gaussian Process?

- A stochastic process is a generalzation of a probability density to functions
- Gaussian processes are stochastic processes that are Gaussian
  - Using Gaussian processes makes computations easier

Why Use Guasssian Process Regression (GPR) ?

- In supervised Machine Learning (ML) we often want to find a function
- GPR focuses on this task
- GPR methods combine
  - Data & Models (casuality)
  - Algorithms and prediction

Supervised Machine Learning and GPR

-Goal: Find a function to predict the data

-What can we do:

1. Select a class of functions, find the best function in that class
2. Test all possible functions, use a prior to weight models

- GPR allows us to try (2)

Supervised Machine Learning and GPR Graphically

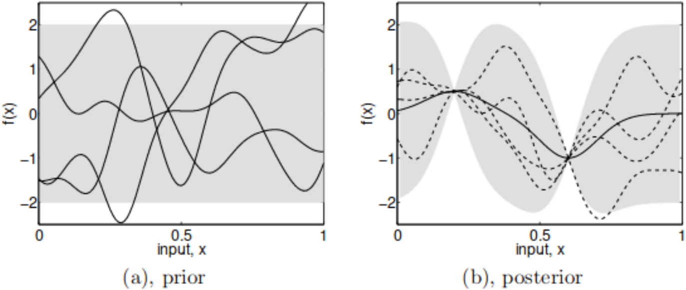

Figure 1: Priors and Posteriors

Basic Regression

- There are several ways to interpret GPR models

1. Function-space view
   - The GP defines a distribution over functions
   - Inference takes place directly in the space of functions
2. Weight-space view
   - More comparable to simple regression methods

The Weight-Space View I

- In a simple linear regression: output is a linear combination of inputs
- In a Bayesian framework we need:
  - A training set D of n observables
- The simple linear model

$$f(x) = x^\top w, \qquad \text{and} \qquad y = f(x) + \varepsilon$$

here,

$$\varepsilon \sim N(0, \sigma_n^2)$$

The Weight-Space View II

- We want to look at the probability density of the observations given parameters

$$
\begin{aligned}
p(y|X,w) &= \prod_{i=1}^{n} p(y_i|x_i,w) \\
&= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}|y - X^\top w|^2 \right) = N(X^\top w, \sigma_n^2 I)
\end{aligned}
$$

- In this Bayesian framework we need to specify a prior on our weights

$$
w \sim N(0, \Sigma_P)
$$

The Weight-Space View III

We have $p(y|X,w)$, knowing this we can apply Bayes' rule to find our posterior

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}, \quad \text{or} \quad p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)}$$

Thus our posterior will be given by

$$p(w|X,y) \propto \exp(-\frac{1}{2\sigma_n^2}(y - X^\top w)^\top (y - X^\top w)) \exp(-\frac{1}{2} w^\top \Sigma_p^{-1} w)$$

Simplifying,

$$p(w|X,y) \sim N\left(\frac{1}{\sigma_n^2} A^{-1} X y, A^{-1}\right), \quad A = \sigma_n^{-2} X X^\top + \Sigma_p^{-1}$$

The Weight-Space View IV

- To make predictions for a test case we average over all possible parameter values
- Our predictive distribution $f_* \equiv f(x_*)$, is given by averaging the output of all possible linear models

$$p(f_*|x_*, X, y) = \int p(f_*|x_*, w)p(w|X, y)dw = \int x_*^\top w \cdot p(w|X, y)dw$$
$$= N\left(\frac{1}{\sigma_n^2} x_*^\top A^{-1} X y, x_*^\top A^{-1} x_*\right)$$

Projecting into a Feature-Space

- We are not limit to linear regression models
- We can replace our linear inputs x with a feature space $\phi(x)$
  - $\phi(x)$ projects x into another space
  - $x : \phi(x) = (1, x, x^2, x^3, \dots)$
- In this case our model is

$$f(x) = \phi(x)^\top w, \qquad \text{and} \qquad y = f(x) + \varepsilon$$

- Our predictive distribution will become,

$$\begin{aligned} p(f_*|x_*, X, y) &= \int p(f_*|x_*, w)p(w|X, y)dw \\ &= \int \phi(x_*)^\top w \cdot p(w|X, y)dw \\ &= N\left(\frac{1}{\sigma_n^2}\phi(x_*)^\top A^{-1}\Phi y, \phi(x_*)^\top A^{-1}\phi(x_*)\right) \end{aligned}$$

The Function-Space View I

- In this setting we use Gaussian Processes to describe a distribution over functions
- Recall: A Gaussian process is a collection of random variables any finite number of which have a joint Gaussian distribution
- By this definition we can completely define a Gaussian process by its mean function and covariance function

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

and

$$K(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$$

.

We will write this Gaussian process as,

$$f(x) \sim GP(m(x), k(x, x'))$$

.

The Function-Space View II

A simple example with Bayesian Linear regression $f(x) = \phi(x)^\top w$ with prior $w \sim N(0, \Sigma_p)$.

Here,

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

and

$$\mathbb{E}[f(x)f(x')] = \phi(x)^\top \mathbb{E}[ww^\top]\phi(x') = \phi(x)^\top \Sigma_p \phi(x')$$

.

We also need a covariance function to specify the covariance between pairs of random variables,

$$\mathrm{cov}(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\left(-\frac{1}{2}|x_p - x_q|^2\right)$$

We can now look at the distribution over functions

$$f_* \sim N(0, K(x_*, x_*))$$

The Function-Space View III

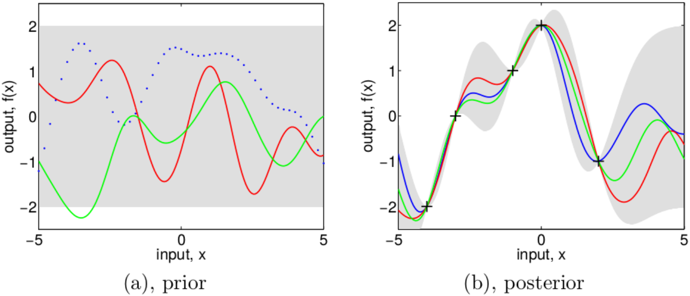

Figure 2: More Priors and Posteriors

Prediction with Noise-Free Observations

The joint distribution of the training outputs f and test outputs f_* is given by,

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = N\left(0, \quad \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

Here,

$$\begin{aligned} f_*|X_*,X,f \sim N(&K(X_*,X)K(X,X)^{-1}f, \\ &K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)). \end{aligned}$$

Prediction with Noisy Observations

Now, $y = f(x) + \varepsilon$

Thus our prior on noisy observations is $\text{cov}(y) = K(X, X) + \sigma_n^2 I$.

The joint distribution of training outputs y and test outputs f_* will be,

$$\begin{bmatrix} y \\ f_* \end{bmatrix} = N\left(0, \quad \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Here,

$$f_* | X_*, X, y \sim N(\bar{f}_*, \text{cov}(f_*))$$

where

$$\bar{f}_* \equiv K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y$$

and

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

A Basic Algorithm for GPR

Using X (inputs), y (targets), k (covariance function), x_* (test input), and σ_n^2 (noise level)

1. L $=$ cholesky$(K + \sigma_n^2 I)$, set $\alpha = L^\top \setminus (L \setminus y)$
2. $\bar{f}_* = k_*^\top \alpha$, set v $= L \setminus k_*$
3. $\mathbb{V}[f_*] = k(x_*, x_*) - v^\top v$
4. $\log p(y|X) = -\frac{1}{2} y^\top \alpha - \sum_i \log L_{i,i} - \frac{n}{2} \log 2\pi$
5. Return, $\bar{f}_*$ (mean), $\mathbb{V}[f_*]$ (variance), and the log marginal likelihood.

Some Code I

How Can We Use GPR in Economics?

- *Optimal Taxation and Insurance using Machine Learning - Sufficient Statistic and Beyond*
  - By Maximilian Kasy
-