THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

Summary: Gaussian processes

Sinead Williamson
The University of Texas Department of Statistics and Data Science

Gaussian distribution

- Perfectly described by its mean and covariance.
- Marginal distribution is Gaussian: If

$$\begin{bmatrix} f \\ g \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right)$$

  then $f \sim \text{Normal}(a, A)$
- Conditional distribution is Gaussian:

$$f|g \sim \text{Normal}(a + CB^{-1}(g - b), A - CB^{-1}C^T)$$

- Conjugate to Gaussian: if $f \sim \text{Normal}(\mu, K)$ and $y|f \sim \text{Normal}(f, \Sigma)$, then

$$f|y \sim \text{Normal}(m, S)$$

  where $S = \left(K^{-1} + \Sigma^{-1}\right)^{-1}$ and $m = S^{-1}(K^{-1}\mu + \Sigma^{-1}y)$

“Infinite-dimensional” Gaussian distribution

- We can think as a function (loosely) as an infinite-dimensional vector f.
- We can then put a distribution over f, to get a distribution over functions.
- We only ever see $f(x)$ at finitely many points $x \in \mathcal{T}$...
- But if our distribution over f is Gaussian, the conditional distribution $p(\{f(x) : x \notin \mathcal{T}\} | f(x) : X \in \mathcal{T})$ is also Gaussian.

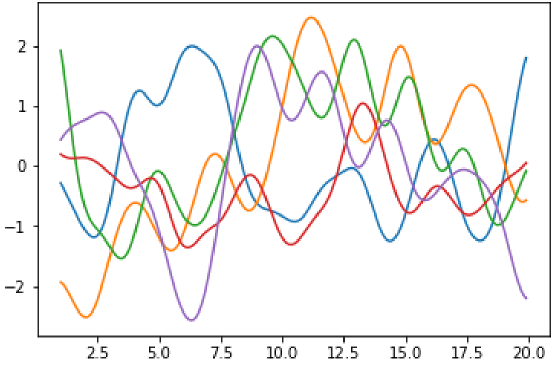
“Infinite-dimensional” Gaussian distribution

- We can think as a function (loosely) as an infinite-dimensional vector f.
- We can then put a distribution over f, to get a distribution over functions.
- We only ever see $f(x)$ at finitely many points $x \in \mathcal{T}$...
- But if our distribution over f is Gaussian, the conditional distribution $p(\{f(x) : x \notin \mathcal{T}\}|f(x) : X \in \mathcal{T})$ is also Gaussian.
- Concretely, we say f is a Gaussian process if all finite marginals are multivariate Gaussian.

Specifying the mean and covariance: Linear regression

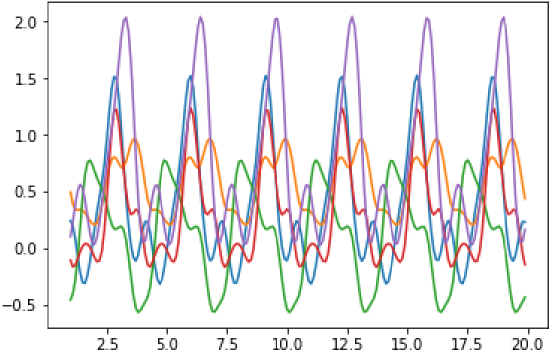
- Everything we've looked at previously falls into this framework!
- Bayesian linear regression: $f(x_i) = \beta^T x_i$, $\beta \sim \mathrm{N}(0, \sigma_\beta^2 I)$...
- So, $f(x_i)$ is normal with covariance $k(i,j) = \sigma_\beta^2 x_i^T x$
- Linear regression is therefore a GP!

Other covariances are more interesting...



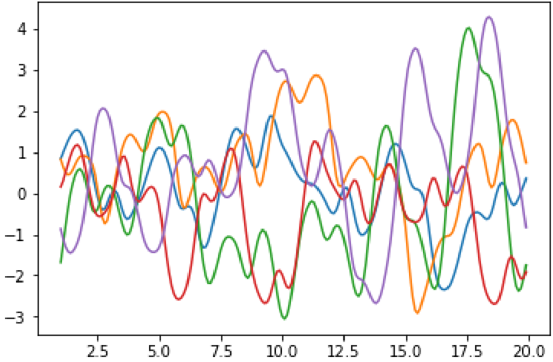
Squared exponential: $k(x, x') = \alpha^2 \exp\left\{-\frac{1}{2\ell^2}(x - x')^2\right\}$

Other covariances are more interesting...



Periodic: $k(x, x') = \alpha^2 \exp\left\{-\frac{2\sin^2((x-x'/p))}{\ell^2}\right\}$

Other covariances are more interesting...



Periodic + squared exponential...

Gaussian process regression

Because of the conditional properties of the Gaussian, we know that:

$$p(f^*|f) = Normal(\tilde{m}, \tilde{K})$$

where

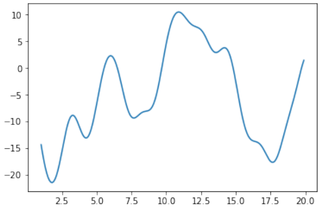
- $\tilde{m} = K(X^*, X)(K(X, X))^{-1}f$
- $\tilde{K} = K(X^*, X^*) - K(X^*, X)(K(X, X))^{-1}K(X, X^*)$

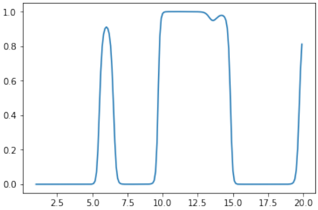
Hyperparameter optimization

- Our kernel will be parametrized by some set of parameters.
- Each parameter setting will give us a different log likelihood.
- We can therefore optimize our hyperparameters to get the best log likelihood!
  - We can easily differentiate our log likelihood to get gradients.
- Alternatively, we can sample hyperparameters in a fully Bayesian scheme.
  - We don't have conjugacy, so we can't Gibbs sample...
  - We can do other things though... Metropolis Hastings is the easiest.
  - Pro: Don't get stuck in local minima, fully explore posterior.
  - Minus: Much slower...

Gaussian process classification

We can do classification with GPs if we transform our function from the reals to the unit interval:





Gaussian process classification

- Let's assume $\pi_i = \Phi(f_i)$, and $y_i \sim \text{Bernoulli}(\pi_i)$
- Equivalently, we can write:
  - $z_i \sim \text{N}(f_i, 1)$
  - $y_i = \begin{cases} 1 & z_i \geq 0 \\ 0 & z_i < 0 \end{cases}$
- If we marginalize out z, this is the same!
- We know $p(z_i | y_i, f)$ is a truncated normal with mean f and variance 1.
- We know $p(f | z_i, x_i)$ is the posterior over a GP, with observations z_i.
- So, we can Gibbs sample from the posterior over f, by alternating samples from f and z.

Gaussian process classification: Logistic variant

- Other choices of squishing function don't have this nice auxiliary variable representation.
- For example, assume $\pi_i = \frac{1}{1+\exp(-f_i)}$.
- Our posterior is proportional to

$$p(f|y,x) \propto N(f;0,K)\prod_i \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- We can approximate this using our Laplace approximation!

Gaussian process classification: Logistic variant

- Other choices of squishing function don't have this nice auxiliary variable representation.
- For example, assume $\pi_i = \frac{1}{1+\exp(-f_i)}$.
- Our posterior is proportional to

$$p(f|y,x) \propto N(f;0,K)\prod_i \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- We can approximate this using our Laplace approximation!
- $L(f) = \log P^*(f|y,x) = \log p(y|f) - \frac{1}{2}f^TK^{-1}f - \frac{1}{2}\log|K|$
- $\nabla L(f) = \nabla \log p(y|f) - K^{-1}f$
- $\nabla\nabla L(g) = \nabla\nabla \log p(y|f) - K^{-1}$
- Approximate posterior with a multivariate normal with precision $\nabla\nabla L(g)$ and mean given by the MAP.

Gaussian process classification: Making predictions

- We have a Gaussian approximation to f at locations x
- We want predictions at locations x^*.
- Let's condition on our MAP approximation for f, and predict at our locations of interest.
- $f^*|f$ is normal, with mean $K(X^*, X)(K(X, X))^{-1}\hat{f}$ and variance $K(X^*, X^*) - K(X^*, X)(K(X, X))^{-1}K(X, X^*)$