Gaussian Process Tutorial

David Jones

Duke University and SAMSI

9th April 2018

Acknowledgements: David Stenning (Imperial College London) contributed some of these slides.
Gaussian Processes in Astronomy

Mention of “Gaussian Process” in SAO/NASA ADS Abstract

Source: http://adsabs.harvard.edu/abstract_service.html
Example: function

\[ y = f(x) \]
Example: no data
Example: function estimation
Example: function estimation

\[ y = f(x) \]
Example: noisy observations
Example: prediction
What is a Gaussian Process?

A GP on the real line is a random real-valued function $f(t)$, which is completely determined by its mean function $m(t)$ and covariance function $C_{tt'} = \text{Cov}(f(t), f(t'))$.

Any finite sample $(f(t_1), \ldots, f(t_n))$ has a multivariate Gaussian distribution with mean $\mu = (m(t_1), \ldots, m(t_n))$ and covariance matrix $\Sigma$, with $\Sigma_{ij} = C_{t_it_j}$

Bivariate Normal Distribution

**Left:** \[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix}
= \begin{pmatrix}
  f(t_1) \\
  f(t_2)
\end{pmatrix} \sim N \left( \bar{\mu} = \begin{pmatrix}
  1 \\
  3
\end{pmatrix}, \Sigma = \begin{pmatrix}
  1 & 0 \\
  0 & 1
\end{pmatrix} \right)
\]

**Right:** \[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix}
= \begin{pmatrix}
  f(t_1) \\
  f(t_2)
\end{pmatrix} \sim N \left( \bar{\mu} = \begin{pmatrix}
  1 \\
  3
\end{pmatrix}, \Sigma = \begin{pmatrix}
  1 & 0.75 \\
  0.75 & 1
\end{pmatrix} \right)
\]
Conditional distributions

$y = f(x)$
Conditional distributions

\[
\begin{pmatrix}
  y_1 \\
  y_2
\end{pmatrix} = \begin{pmatrix}
  f(t_1) \\
  f(t_2)
\end{pmatrix}
\]

\[y_2 | y_1 = -1.52 \sim N(-1.2, 0.62^2)\]
Suppose $\vec{y}_1$ are values we observe, and $\vec{y}_2$ are values we want to predict, then:

$$(\vec{y}_1, \vec{y}_2) \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$\vec{y}_2 \mid \vec{y}_1 \sim \mathcal{N} \left( \Sigma_{21} \Sigma_{11}^{-1} \vec{y}_1, \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right)$
\[
\begin{pmatrix}
y_1 \\
y_2 
\end{pmatrix} = \begin{pmatrix} f(-5) \\ f(-4) \end{pmatrix} \sim N \left( \vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.790 \\ 0.790 & 1 \end{pmatrix} \right)
\]

Applying the conditional Gaussian result

\[
y_2 | y_1 = -1.52 \sim N \left( 0.790(1)^{-1}(-1.52), 1 - 0.790(1)^{-1}0.790 \right)
\]

\[
y_2 | y_1 = -1.52 \sim N(-1.2, 0.62^2)
\]
Gaussian Process

- Assume $y = f(x)$ is a univariate function of d-dimensional $x$
- For a zero-mean Gaussian Process (GP), any (finite) collection $y_1, \ldots, y_m$ corresponding to $x_1, \ldots, x_m$ is distributed

$$\vec{y} \sim \mathcal{N}(\vec{0}, \Sigma)$$

where $\Sigma_{ij} = R(x_i, x_j)$
- $R(x, x')$ is a covariance function (i.e. kernel) that we specify.
  - A common choice is the squared exponential kernel:

$$R_{SE}(x, x') = \sigma^2 \exp \left( -\frac{(x - x')^2}{2l^2} \right)$$

- $\sigma^2$ is a scale factor (all kernels have this term)
- The length-scale, $l$, controls the "wiggliness" of the function
Gaussian Process

- Assume \( y = f(x) \) is a univariate function of d-dimensional \( x \)
- For a zero-mean Gaussian Process (GP), any (finite) collection \( y_1, \ldots, y_m \) corresponding to \( x_1, \ldots, x_m \) is distributed

\[
\tilde{y} \sim \mathcal{N}\left(\mathbf{0}, \Sigma\right)
\]

where \( \Sigma_{ij} = R(x_i, x_j) \)

- \( R(x, x') \) is a covariance function (i.e. kernel) that we specify.
  - A common choice is the squared exponential kernel:

\[
R_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)
\]

- \( \sigma^2 \) is a scale factor (all kernels have this term)
- The length-scale, \( l \), controls the “wiggliness” of the function
Periodic and Locally Periodic Kernels

- A periodic kernel models functions that repeat (periodically):

\[ R_{\text{Per}}(x, x') = \sigma^2 \exp \left( -\frac{2\sin^2(\pi|x - x'|/p)}{l_p^2} \right) \]

- A locally periodic kernel yields functions with a periodic component that may evolve over time:

\[ R_{\text{LocPer}}(x, x') = \sigma^2 \exp \left( -\frac{2\sin^2(\pi|x - x'|/p)}{l_p^2} \right) \exp \left( -\frac{(x - x')^2}{2l_e^2} \right) \]

- A good resource: The Kernel Cookbook (by David Duvenaud)
Another View of Kernels

squared exponential with $\sigma^2 = 1$ and $l = 1$

periodic with $\sigma^2 = 1$, $l = 1$, and $p = 2$
GP Draws: Squared Exponential Kernel

squared exponential with $\sigma^2=1$ and $l=1$

squared exponential with $\sigma^2=1$ and $l=2$

squared exponential with $\sigma^2=1$ and $l=0.5$
Inference with Gaussian Processes

Let $\vec{y}_1$ be some values we observe and $\vec{y}_2$ are values we want to predict. Then:

$$
\begin{pmatrix}
\vec{y}_1 \\
\vec{y}_2
\end{pmatrix} \sim N\left(\begin{pmatrix}
\vec{0} \\
\vec{0}
\end{pmatrix}, \begin{pmatrix}
R_{11} & R_{12} \\
R_{21} & R_{22}
\end{pmatrix}\right)
$$

$$
\vec{y}_1 \mid \vec{y}_2 \sim N\left(R_{11} R_{22}^{-1} \vec{y}_2, R_{11} - R_{12} R_{22}^{-1} R_{21}\right)
$$

- Mean for the new points is a weighted average of the observed points.
- Mean of a new point approaches value of an observed point as the new point approaches the observed point.
- Variance of a new point goes to zero as the new point approaches an observed point.
Inference with Gaussian Processes

Unconditional GP Draws

Observed Data

Conditional GP Draws
What are we actually doing?

- When using GPs, we are specifying a prior on the relationship between $t$ and $f(t)$, instead of some parameters that describe this relationship
  - i.e. “nonparametric”
- GPs especially useful for prediction; (maybe) not as useful for making inference about the relationship
  - e.g., useful for predicting sunspot cycle; less useful for learning about the cycle
Noisy observations
Gaussian Processes in the case of noisy observations

- Now $y_1, \ldots, y_m$ corresponding to $x_1, \ldots, x_m$ is distributed

$$\vec{y} \sim \mathcal{N}(\vec{0}, \Sigma + \tau^2 I_m)$$

where $\Sigma_{ij} = R(x_i, x_j)$.

- More specifically the model is

$$\vec{y} \sim \mathcal{N}(\vec{f}, \tau^2 I_m)$$

$$\vec{f} \sim \mathcal{N}(\vec{0}, \Sigma)$$

where $\vec{f} = (f(x_1), \ldots, f(x_m))^T$
Hyper-parameters

- For real research problems, we often (always?) lack the information needed to fix the parameters of the covariance function
- Typical solutions:
  - Maximum likelihood estimation
  - Cross validation
  - Specify some prior distributions and do MCMC
- Caveat: $C^{-1}$ is $O(N^3)$; exploit sparsity if possible
Underlying Model + Correlated Noise

Image: https://astrobites.org/2014/07/01/beyond-chi-squared-an-introduction-to-correlated-noise/
Consider the following setup:

- Physical model: \( g_\phi(t) = a_1 \sqrt{10t} + a_2 \sqrt{10t} \exp\left(\frac{-10t}{a_3}\right) \)
- Physical parameters: \( \phi = (a_1, a_2, a_3) \)
- Reality: \( a_1 = 1, \ a_2 = 0.5, \ a_3 = 2 \)

- Have 11 observations with correlated noise
- We want to infer \( a_1, a_2, \) and \( a_3 \)
Toy Example: observations

![Graph showing the true mean curve with observations plotted against time (t) and the response variable (y).]
Toy Example: model formulation

Covariance Function (Kernel):

$R(t, t') = \sigma^2 \exp\left(-\beta (t - t')^2\right) + \delta_{tt'} \tau^2$

$\delta_{tt'} = 1$ if $t = t'$ and 0 otherwise

Sampling Model:

$\tilde{y} \sim N(g_\phi(\tilde{t}), \Sigma)$

$\Sigma_{ij} = R(t_i, t_j)$

$\phi = (a_1, a_2, a_3)$

Priors:

$\beta \sim \text{Exponential}(1)$

$\sigma^2 \sim \text{Inv-Gamma}(5, 0.1)$

$\tau^2 \sim \text{Inv-Gamma}(5, 0.01)$

Flat priors on $a_1$, $a_2$, and $a_3$

Model Fitting:

Parameters estimated with one-at-a-time Metropolis MCMC
Toy Example: results
Toy Example: results

-3 -2 -1 0 1 2

0.0 0.2 0.4 0.6 0.8 1.0

-3 -2 -1 0 1 2

true mean curve
predicted mean curve
95% credible interval
Real Example I: Czekala et al. 2015

Figure 11. The $K$-band SPEX spectrum of Gl 51 (blue) compared with a PHOENIX model (red) generated by drawing parameters from the inferred posterior distribution. (bottom) The residual spectrum along with contours representing the distributions of a large number of random draws from the covariance matrix (the shading is representative of the 1, 2, and 3σ spreads of that distribution of draws), as in Fig. 9. Note how the ‘outlier’ features (Na I at 2.21 μm and Ca I at 2.26 μm) are identified and treated by the local covariance kernels.

- Likelihood framework for spectroscopic inference based on synthetic model spectra and GPs
- Addresses mismatches in model spectral line strengths w.r.t. data due to intrinsic model imperfections
Example II: Mars Rover ChemCam

Artistic rendering of ChemCam LIBS analyses using NASA’s Mars Curiosity Rover
**General concept:** Estimate the settings of a theoretical model’s input parameters $\theta$ that are consistent with physical measurements $y$.

$$y = \eta(\theta) + \delta + \varepsilon$$

Measured and modeled LIBS spectra of basalt.

Slide courtesy Kary Myers (LANL)
Model the relationships between the apparent RV of a star due to a spot and proxies for stellar variability

Use locally periodic kernel

\[ R_{\text{LocPer}}(t, t') = \sigma^2 \exp \left( -\frac{2\sin^2(\pi|t - t'|/p)}{l_p^2} \right) \exp \left( -\frac{(t - t')^2}{2l_e^2} \right) \]
GPs in Python

Packages include:

- scikit-learn
- GPflow
- PyMC3
- George
- ...

Many good tutorials online e.g.

For more information...

- Contact me:
  - dav.jones2000@gmail.com
  - dej17@duke.edu