11/1/19

# 8803 Learning from Data: Lecture 20

Notebooks for today (in gaussian-processes subdirectory):
demo-GaussianProcesses.ipynb
Gaussian_processes_exercises.ipynb

Websites: "Gaussian process webapps" page on Carmen
(A) "Sample Size Calculations for Computer Experiments"
(B) "Gaussian process regression: a function space perspective"
(C) "Gaussian process regression demo"

- Finish going through demo-GaussianProcesses.ipynb

- Games with (A), (B), (C)
  - Start with (A) and "Sample Path Plots"
    - Try changing "Correlation length" (switch "No. of realizations" for new draws)
    - Change "Select correlation family"
      - Note extra parameter with Matérn
    - See "About" tab for formulas.

  - For (B), you should see successive draws from a GP
    - Try changing covariance and length scale; does it change as predicted?
    - Now try adding points. What happens?
    - Try adjusting noise. What happens?

  - For (C)
    - Click "Show mean and credible intervals" and "Sample independently"
    - add observations
    - add a new process.

11/1/19

- Brief introduction to GPs from Melendez et al., Phys Rev C 100, 044001 (2019).

- Gaussian processes (GPs) are often used for nonparametric regression. Cf. fitting a polynomial, where the basis functions are $1, x, x^2, \ldots, x^d$ and the coefficients are parameters. This is parametric regression.

- The most common applications of GPs are interpolation and regression. To carry these out we need to calibrate (eg. fit) the GP parameters.

- A GP is specified by a mean function $m(x)$ [often written $\mu(x)$] and a positive semidefinite covariance function, or kernel, $K(x,x')$ where $x \in \mathbb{R}^d$ and $d$ is the dimension of parameter space.
  - If we know $m(x)$ and $K(x,x')$, then a GP $f(x)$ is denoted
$$f(x) \sim GP[m(x), K(x,x')]$$
  in the same way a normal distribution for $g(x) \sim N(\mu, \sigma^2)$

[we express the $d$ dependence here]

- While the definition has continuous, infinite-dimensional $x$, in practice we use a finite number of points.
  $\vec{x} = \{x_i\}_{i=1}^N \leftarrow N$ input points (vectors actually) at $\vec{f} = \{f(x_i)\}_{i=1}^N$

Define $\vec{m} = m(\vec{x}) \in \mathbb{R}^N$ and $K = K(\vec{x}, \vec{x}) \in \mathbb{R}^{N \times N}$

$$\Rightarrow \vec{f} | \vec{x} \sim N(\vec{m}, K)$$

"$\vec{f}$ conditional on $\vec{x}$" ← definition of GP: any subset of inputs are a multivariate Gaussian

- The mean function is the a priori "best guess" of $f$. If no features, often taken as 0.

- Our specification of the kernel tells us what $K$ is.

11/1/19

- So how do we use this GP? Let's assume we already know $\vec{\theta}$, the set of hyperparameters. And suppose we know the value of the function $\vec{f}$ at a set of $\vec{x}_1$ points ⇒ this is our <u>training set</u>

⇒ partition $\vec{x} = [\vec{x}_1 \; \vec{x}_2]^T$ and $\vec{f} = f(\vec{x}) = [\vec{f}_1 \; \vec{f}_2]^T$

into $N_1$ training and $N_2$ test points (the latter are our predictions).
- There are corresponding vectors $\vec{m}_1, \vec{m}_2$ and covariance matrices.

The <u>definition</u> of a GP says the joint distribution of $\vec{f}_1, \vec{f}_2$ is

$$\begin{bmatrix}\vec{f}_1 \\ \vec{f}_2\end{bmatrix} \Big| \vec{x}, \vec{\theta} \sim N\left(\begin{bmatrix}m_1 \\ m_2\end{bmatrix}, \begin{bmatrix}K_{11} & K_{12} \\ K_{21} & K_{22}\end{bmatrix}\right)$$

where $K_{11} = \kappa(\vec{x}_1, \vec{x}_1)$, $K_{22} = \kappa(\vec{x}_2, \vec{x}_2)$, $K_{12} = \kappa(\vec{x}_1, \vec{x}_2) = K_{21}^T$

- Then manipulation of matrices tells us

$$\vec{f}_2 | \vec{x}_2, \vec{f}_1, \vec{\theta} \sim N(\tilde{m}_2, \tilde{K}_{22})$$

This makes sense:
$p(\vec{f}_2, \vec{f}_1) = p(\vec{f}_2 | \vec{f}_1) \times p(\vec{f}_1)$
$\Rightarrow p(\vec{f}_2 | \vec{f}_1) = p(\vec{f}_2, \vec{f}_1) / p(\vec{f}_1)$
and the ratio of gaussians is also a gaussian ⇒ just need to figure out the matrices!

where

$$\tilde{m}_2 = m_2 + K_{21} K_{11}^{-1} (\vec{f}_1 - \vec{m}_1)$$ ← so this is our best prediction (solid line)

$$\tilde{K}_{22} \equiv K_{22} - K_{21} K_{11}^{-1} K_{12}$$ ← this is the variance (determines band)

- We can make draws from $N(\tilde{m}_2, \tilde{K}_{22})$ and plot. With dense enough grid, we will have lines that go through the $\vec{x}_1$ points and have bands in between, which grow larger

Gaussian white — if same, otherwise ↓ diagonal

- If we have noise at each point, then $K_{11} \to K_{11} + \sigma_n^2 I_{N_1}$. Need to add this, even if perfect data, for numerical reasons: $K_{11}^{-1}$ may be unstable without it (called adding a "nugget").
- For predictions with noise, then $K_{22} \to K_{22} + \sigma_n^2 I_{N_2}$

11/1/19

Aside: How do we make draws from a multivariate Gaussian (normal) distribution if we know how to make draws from $N(0,1)$, a standard distribution.

- Suppose $\vec{x} \sim N(\vec{\mu}, \Sigma)$, so $p(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{\sqrt{\det \Sigma (2\pi)^d}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$
  ($\vec{x}$: dimension $d$; $\Sigma$: same dim $d$)
- Generate $\vec{u} \sim N(0, I_d)$ (from $d$ draws of $N(0,1) \Rightarrow \vec{u} = \begin{pmatrix} u_0 \\ \vdots \\ u_{d-1} \end{pmatrix}$)
- Find $L$ such that $\Sigma = LL^T$ (Cholesky decomposition)
- Then $\vec{x} = \vec{\mu} + L\vec{u}$ has the desired distribution.

Check: What is expectation value of $\vec{x}$, given $E[\vec{u}] = 0$?

$E[\vec{x}] = E[\vec{\mu} + L\vec{u}] = \vec{\mu} + L E[\vec{u}]^0$ since linear returns the mean.
$= \vec{\mu}$ ✓

Now try (variance of $\vec{u}$)

$\Sigma \stackrel{?}{=} E[(\vec{x}-\vec{\mu})(\vec{x}^T-\vec{\mu}^T)] = E[(L\vec{u})(L\vec{u})^T] = E[L\vec{u}\vec{u}^T L^T] = L E[\vec{u}\vec{u}^T] L^T$
$= L I_d L^T = \Sigma$ ✓

So it works!

---

Let's check that it works for the special case of one training, one test point.

- Let's also take mean zero to avoid clutter.
- Then $\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \equiv \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ so Gaussian with

claim $f_2|f_1 \sim N\left(\Sigma_{21}\Sigma_{11}^{-1} f_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$ mean $\Sigma_{21}\Sigma_{11}^{-1} f_1$ and variance $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Try it explicitly

$p(f_2|f_1) = p(f_1,f_2)/p(f_1) = \frac{\frac{1}{\sqrt{\det\Sigma (2\pi)^2}} e^{-(f_1 f_2)\Sigma^{-1}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}}}{\frac{1}{\sqrt{2\pi\Sigma_{11}}} e^{-f_1^2\Sigma_{11}^{-1}}}$ All of the $\Sigma_{ij}$'s are just scalars now and $\Sigma_{12} = \Sigma_{21}$

$\Sigma^{-1} = \frac{1}{(\Sigma_{11}\Sigma_{22} - \Sigma_{12}\Sigma_{21})} \begin{pmatrix} \Sigma_{22} & -\Sigma_{21} \\ -\Sigma_{12} & \Sigma_{11} \end{pmatrix} \Rightarrow (f_1 f_2)\Sigma^{-1}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = (-\Sigma_{22} f_1^2 + 2\Sigma_{21} f_1 f_2 - \Sigma_{11} f_2^2)/\det\Sigma$

($\det\Sigma = \Sigma_{11}\Sigma_{22} - \Sigma_{12}\Sigma_{21}$)

So if we gather the $\Sigma_{22}$ parts, we find to complete the square we need $(f_2 - \Sigma_{21}\Sigma_{11}^{-1} f_1)$, which identifies the mean, and then the variance part comes out as advertised.