Gaussian process regression (1)

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

November 4, 2020

Outline

Previously:

- Covariance between parameters
- Interactions between predictors

Today:

- Gentle intro to Gaussian processes

What is a GP?

Regression with normal likelihood

Yet again we turn back to our old friend the linear model:

$$y_i \sim \mathrm{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x$$

(priors omitted)

What we're saying: y_i normally distributed around a regression line $\mu_i = \alpha + \beta x$.

But of course, there is no reason why this has to be a line.

Regression with normal likelihood

Regression around an arbitrary function:

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= f(x)
\end{aligned}
$$

(priors omitted)

The design and choice of f makes up the flavor of the regression.
f could be:

- a function derived from a scientific model
- a sum of basis functions with parametric weights
- something else

GP regression offers an option for the “something else.”

GP: the definition

A Gaussian process is a random *function* – i.e., we're really talking about a probability distribution on a space of functions.

The feature that makes a GP a GP: if you pick any n values of x, then the vector of function values $(\mu(x_1), \mu_(x_2), \ldots, \mu(x_n))$ has a multivariate normal distribution:

$$(\mu(x_1), \ldots \mu_(x_n)) \sim \mathrm{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n))$$

The GP is determined by its mean function m and covariance K.

GP: the definition

Typically, the covariance matrix is determined by a function called the *kernel* $k(x, x')$.

- $k(x, x')$ determines how much the value of $\mu(x)$ depends on $\mu(x')$.
- Common (not universal) property: $k(x, x')$ depends on the distance between x, x'
- Idea: we're looking for continuous functions, so the values of $\mu(x), \mu(x')$ should be close if x, x' are close; but if they're far apart

Squared exponential covariance

Very common choice: squared exponential covariance function:

$$k(x, x') = \eta^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

Covariance is high when $x - x'$ is small, falls off at longer ranges.

Hyperparameters:

- η: the maximum covariance
- ℓ: the *length scale*, controls how quickly covariance decays.

Let's explore the behavior of this and other GPs.

Example: tool complexity in Polynesian islands

Tool complexity

Data: population and tool complexity among ancient Polynesian island cultures

Hypothesis: number of distinct tools found on an island is a function of (log) population

- More people $\rightarrow$ more invention
- Diminishing returns as population increases

The model

Simple model: Poisson GLM

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \beta \log P_i \\
\alpha &\sim \mathrm{Normal}(0, 10) \\
\beta &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

But we want to add something to the model to account for trade. Islands can acquire tools without inventing them if they have contact with other islands.

The data

The data contain a column classifying the islands as high contact / low contact. So we could fit a simple varying-intercepts model:

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \gamma_{C[i]} \beta \log P_i \\
\alpha &\sim \mathrm{Normal}(0, 10) \\
\gamma_{C[j]} &\sim \mathrm{Normal}(0, 1) \\
\beta &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

where $C[i] \in \{\mathrm{high}, \mathrm{low}\}$.

Better version

But, we can make a finer-grained model. What determines whether and how frequently two islands can trade?

Better version

But, we can make a finer-grained model. What determines whether and how frequently two islands can trade?

One option: distance

- Add a varying-intercepts term that is a function of distance
- Acts like a categorical intercept, but dependent on a continuous predictor
- Use squared-exponential kernel for covariance

GP version

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \gamma_i \beta \log P_i \\
\alpha &\sim \mathrm{Normal}(0, 10) \\
\beta &\sim \mathrm{Normal}(0, 1) \\
\gamma_i &\sim \mathrm{MVNormal}(0, \mathsf{K}) \\
\mathsf{K}_{ij} &= \eta^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \\
\eta &\sim \mathrm{HalfCauchy}(1) \\
\ell &\sim \mathrm{HalfCauchy}(1)
\end{aligned}
$$

where the covariances K_{ij} are computed using the squared-exponential kernel.

Why is this a GP?

This GP looks and feels a bit different from the examples before.

- Not explicitly computing a regression across many values of x
- Still just fitting one varying intercept per observation

But: really, our model looks like this:

$$
\begin{aligned}
y_i &\sim \mathrm{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \gamma(\mathrm{lat}, \mathrm{long})\beta \log P_i \\
\alpha &\sim \mathrm{Normal}(0, 10) \\
\beta &\sim \mathrm{Normal}(0, 1) \\
\gamma_{lat,long} &\sim \mathcal{GP}(\prime, \|) \\
k(\mathrm{x}, \mathrm{x}') = \eta^2 \exp\left(-\frac{\|\mathrm{x} - \mathrm{x}'\|}{2\ell^2}\right) \\
\eta &\sim \mathrm{HalfCauchy}(1)
\end{aligned}
$$

Summary

Gaussian processes allow highly flexible function fitting

Next week:

- More Gaussian process regression; computational issues