CS6190: Probabilistic Modeling **Due Th 4/28**

Homework 4: Gaussian Processes

Instructions: Be sure to electronically submit your answers in either R Markdown (*.Rmd) or Sweave (*.Rnw) format. Include all of the output of your code, plots, and discussion of the results in your write up. You may work together and discuss the problems with your classmates, but write up your final answers entirely on your own.

1. Implement Gaussian process regression for univariate variables y and x:

$$\begin{aligned} y &\sim f(x) + \epsilon, \\ f &\sim \mathrm{GP}(0, k), \\ \epsilon &\sim N(0, \sigma^2). \end{aligned}$$

   Use a squared exponential kernel:

$$k(x, x') = \exp\left(-\frac{1}{2\lambda}(x - x')^2\right),$$

   where λ is a parameter controlling the kernel width.

   Draw some plots of curves sampled from just the prior distribution. Try a handful of different settings for λ to see what effect is has. NOTE: you will need to add a small value to the diagonal of your covariance matrix to ensure it is invertible.

2. Test your Gaussian process regression with the following example. Generate synthetic data from the model:

$$\begin{aligned} y &= \sin(x) + \epsilon, \\ \epsilon &\sim \mathrm{N}(0, \sigma^2). \end{aligned}$$

   Start by generating x values randomly uniform on the interval $[0, 2\pi]$. Then generate your y values using $\sigma = 0.5$ and a sample size of $n = 50$.

   Now, plot (1) your raw data, (2) the true answer, (3) your estimated posterior mean function from Gaussian regression, and finally (4) the 95% confidence region for your posterior mean. (You might refer to the file `BayesianLinearRegression.r` for examples of this kind of plotting.) What value of λ did you find worked well?

3. Download the data `SaltLakeTemperatures.csv` from the class website. Let's use Gaussian process regression to model this data. This is historic monthly temperature data from a weather station in Salt Lake City for dates ranging from 1905 - 2015. This data is from the National Ocean and Atmospheric Administration (NOAA)[1]. The column `AVEMAX` is the average daily maximum temperature within the month of measurement. First, you will need to create two more columns in the data frame: one for the `MONTH` (as a number 1 - 12), and one for the `YEAR` (as a number 1905 - 2015). This should be extracted from the `DATE` column, which is an 8-digit number in the format `YYYYMMDD`.

   Now try the following:

[1] `http://www.ncdc.noaa.gov/cdo-web/#t=secondTabLink`

(a) Model the average cyclical temperature trajectory over the course of a year. In other words, use your `MONTH` column as the x variable. Again, plot (1) the raw data, (2) your average trajectory, and (3) the 95% credible region.

(b) Next, say you want to model the historic trend in this cyclical temperature trajectory. That is, how has this annual temperature cycle changed over the years? Use the following hierarchical model:

$$\begin{aligned} y &= f_0(x) + f_1(x)z + \epsilon, \\ f_0(x) &\sim \mathrm{GP}(0, k(x, x')), \\ f_1(x) &\sim \mathrm{GP}(0, k(x, x')), \\ \epsilon &\sim \mathrm{N}(0, \sigma^2). \end{aligned}$$

Here the x variable is again the `MONTH` data, and now the z variable is your `YEAR` data. So, what is going on here is the following: $f_0(x)$ is a smooth function representing the "intercept" annual temperature cycle, i.e., the temperature cycle in 1906, and $f_1(x)$ is the "slope", which means a linear historical trend that is defined at each time during the year. You should subtract 1906 from your `YEAR` data to make $f_0(x)$ be the baseline curve at 1906.

The big question at the end is in the posterior distribution of the $f_1(x)$ function—it tells us if temperatures in SLC are getting hotter, colder, or staying the same, and it tells us this as a function of the time of year. In the end, what conclusions can you draw? (Again, it will be important to think about parameter selection and confidence intervals, etc.)

(c) **Optional: if you want to go beyond what's required.** Try using Type II Maximum Likelihood to estimate the best λ and σ parameters from the data.