Supervision Assignments in MLBI
Easter Term 2019
Set 4

Supervisor : Dionysis Manousakas
dm754@cam.ac.uk

April 21, 2019

1. *(Bayesian linear and Gaussian process regression.)*

   Plot the time series of monthly mean global CO2 concentrations obtained from the file `co2.txt` (original data obtained from `http://www.esrl.noaa.gov/gmd/ccgg/trends`).

   We will apply Bayesian linear and Gaussian process regression to predict the CO_2 concentration $f(t)$ as a function of time t, where $t = \text{Year} + (\text{Month} - 1) = 12$.

   (a) First we model the function using linear regression, that is, using the functional form

   $$f(t) = at + b + \epsilon(t)$$

   with i.i.d. noise residual $\epsilon(t) \sim N(0, 1)$ and prior $a \sim N(0, 10^2)$, $b \sim N(360, 100^2)$. Compute (e.g. using MATLAB) the posterior mean and covariance over a and b given the CO_2 data.

   (b) Let a_{MAP}, b_{MAP} be the MAP estimate in the question above. The residual is the difference between the observed function values and the predicted mean function values

   $$g_{\text{obs}}(t) = f_{\text{obs}}(t) - (a_{\text{MAP}}t + b_{\text{MAP}}),$$

   where $f_{\text{obs}}(t)$ is the observed value of the CO_2 concentration at time t. Plot $g_{\text{obs}}(t)$. Do you think these residuals conform to our prior over $\epsilon(t)$? State, with justifications, which characteristics of the residual you think do or do not conform to our prior belief.

   (c) Write a MATLAB function to generate samples drawn from a GP. Specifically, given a covariance kernel function $k(\cdot, \cdot)$ and a vector of input points $\mathbf{x}$, return a function $f(\mathbf{x})$ evaluated on the input points $\mathbf{x}$ drawn randomly from a GP with the given covariance kernel and with zero mean.

   (d) Test your function by plotting sample functions drawn from the following kernel, for various settings of the hyperparameters

   $$k(s,t) = \theta^2 \Big(\exp\Big(-\frac{2\sin^2(\pi(s-t)/\tau)}{\sigma^2} \Big) + \phi^2 \exp\Big(-\frac{(s-t)^2}{2\eta^2} \Big) \Big) + \zeta^2 \delta_{s=t}$$

   Describe the characteristics of the drawn functions, and how the characteristics of the functions depend on the parameters.

   (e) Suppose we were to consider modelling the residual function $g(t)$ using a zero mean GP with the covariance kernel above. Based on the plot of $g(t)$ and your explorations in the preceding part, what do you think will be suitable values for the hyperparameters of k?

(f) Extrapolate the CO_2 concentration levels to 2020 using the GP with covariance kernel k of the equation in (d), and your chosen parameter values. Specifically, compute the predictive mean and variance of the residual $g(t)$ for every month between September 2007 and December 2020 given the observed residuals $g_{\text{obs}}(t)$. Plot the means and one standard deviation error bars of the extrapolated CO_2 concentration levels

$$f(t) = a_{\text{MAP}}t + b_{\text{MAP}} + g(t)$$

along with the observed CO_2 levels. Does the behaviour of the extrapolation conform to your expectations? How sensitive are your conclusions to settings of the kernel hyperparameters?

(g) Why is the above procedure not Bayesian? How would we go about modelling $f(t)$ in a Bayesian framework?