DRAFT, 4/17/2020

CSCE 689-606 Spring 2020

Major Project

Due: 11:59pm Monday, May 4, 2020

In prior assignments, you developed parallel implementations of the Gaussian Process Regression (GPR) technique to predict the value of a function at a point in a two-dimensional unit square using known values of the function at points on a grid laid out on the unit square. The GPR model was defined by hyper-parameters that were provided to you.

In this project, we will explore how to compute the hyper-parameters that can be used in the GPR model to predict values with high accuracy. The prediction f_* at a point *q(x,y)* is given as

$$f_* = k_*^T (tI + K)^{-1} f \tag{1}$$

where *K* is the kernel matrix that represents correlation between *f* at grid points. Specifically, for two points *r(x,y)* and *s(u,v)*:

$$K(r,s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2l_1^2} + \frac{(y-v)^2}{2l_2^2}} \tag{2}$$

in which , *l1* and *l2* are two hyper-parameters that should be chosen to maximize the likelihood of the prediction being accurate. The vector *f* denotes the observed data values at the grid points. The vector k_* is computed as given below:

$$k_*(q,s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2l_1^2} + \frac{(y-v)^2}{2l_2^2}}, \tag{3}$$

for all grid points *s*.

To estimate *l1* and *l2* we split the data into two sets randomly: 90% of the points form the training set and the remaining 10% form the test set. We select initial values for the parameters *l1* and *l2* and construct *K* using points in the training set. Next, we predict at each test point using Eq. (1). Using predictions at all the test points, we compute the mean square error of the predictions from the observed data:

$$mse = \sum_{i=1}^{n_t} \left(f_*(r_i) - f(r_i)\right)^2, \tag{4}$$

where *nt* is the number of test points. Our goal is to determine those values of *l1* and *l2* that minimize mse. There are a number of approaches to explore the hyper-parameter space. We will use the grid search technique where we evaluate mse at grid points in the hyper-parameter space and select the hyper-parameter values that result in the smallest mse. For example, if we anticipate the hyper-parameters to lie in the interval [0.1,1], we can assign *l1* and *l2* values from 01. to 1 in increments of 0.1. For two parameters, there will be 100 distinct pairs. For each assignments, we will compute mse using Eq. (3) which requires compute predictions at each test point using the kernel function shown in (2) that depends on *l1* and *l2*.

Matlab files that implement the algorithm are provided as an illustration.

1. (75 points) In this assignment, you have to develop parallel code to determine the hyper-parameters *l1* and *l2* that minimize the mse. You can design an OpenMP-based shared memory code or a GPU code for this project. You are encouraged to modify the code you developed in earlier assignments for Eq. (1).

2. (20 points) Describe your strategy to parallelize the algorithm. Discuss any design choices you made to improve the parallel performance of the code. Also report on the parallel performance of your code.
3. (5 points) Apply your code to another data set to see how it performs. You may choose any appropriate data set that you can find.

Submission: You need to upload the following to eCampus:

1. Submit the code you developed.
2. Submit a single PDF or MSWord document that includes the following.
    - Responses to Problem 1, 2, and 3. Response to 1 should consist of a brief description of how to compile and execute the code on the parallel computer

Helpful Information:

1. Source file(s) are available on the shared Google Drive for the class.