Option 6: Gaussian Process Regression

In this project, you have to implement a Gaussian Process Regression model. “Gaussian processes are a powerful, non-parametric tool that can be used in supervised learning, namely in regression but also in classification problems” [1]. To understand the basic concepts behind Gaussian Processes we recommend you to read [1] or any other tutorial/book you may find useful. However, to describe this project, we will refer to the equations in [1].

You will be given a training set containing feature vectors and the outputs which are real numbers. For an unseen value in the test set, you have to predict the output value. This is the setting of a regression problem. You have to find the regression function using the paradigm of Gaussian processes.

Your first task: Find the output (or the regressed value) for a given input vector, as shown in Equation 9. Specifically, you have to get $\overline{f_*}$. As you can see, you need values for various parameters, namely, σ_f, l (in Eqn 3), σ_n (in Eqn 7). You can assume any arbitrary values for the above parameters and report the obtained output values for each of the given input vector and the accuracy. The accuracy can be calculated as root mean square error between the true output (on training set) and your obtained output. *Note: For this task, you can report any output you get. It is not important what value of the output you get.*

Your second task: Instead of setting the values arbitrarily, you will have to infer l, σ_f, σ_n as shown in section 4 by minimizing the negative log-likelihood of the conditional distribution $p(y|x)$. Derive this log-likelihood and prove that it is as shown in Equation 12. Compute its derivative w.r.t. the parameters and prove that it is equal to the term shown in Equation 14. You may need to use some results regarding matrix derivatives. Minimize the negative log-likelihood w.r.t the parameters (i.e. $\theta_k = \{l, \sigma_f, \sigma_n\}$) as shown in (14) and (15). You can use simple gradient descent. However, you are free to use any optimization algorithm / library. Once you get the parameters i.e. $\{l, \sigma_f, \sigma_n\}$, use them to compute the regressed values for the same inputs used in the first task. Report the change in accuracy with the same inputs you used in the first task.

Evaluation: We will withhold some output for test data. The quality of your regressor output will be matched against the true output using the RMSE metric.

Reference:

[1] Melo, José. "Gaussian Processes for regression: a tutorial."