Critical Review, Nate Schambach, Team 10 – Optimized Tissue Reconstruction

Paper: Gaussian Process Regression: Active Data Selection and Test Point Rejection

Seo, S., Wallat, M., Graepel, T., Obermayer, K., *Gaussian Process Regression: Active Data Selection and Test Point Rejection.* Department of Computer Science, Technical University of Berlin, 2000.

General Overview Of Gaussian Process:

The paper does a very succinct and clear description of a Gaussian process so everything here follows directly from there.

Gaussian Process Regression is a regression method which utilizes a collection of random variables that have joint Gaussian distributions. This joint distribution is:

$$P(\boldsymbol{t}|\boldsymbol{C}, \boldsymbol{x_n}) = \frac{\mathbf{1}}{Z}\exp(-\frac{1}{2}(\boldsymbol{t}-\mu)^T\boldsymbol{C}^{-1}(\boldsymbol{t}-\mu))$$

Where $\boldsymbol{C}$ is the covariance matrix defined by the covariance function C and μ is the mean function. The regression makes a prediction utilizing these as such:

$$\hat{y}(\tilde{x}) = \boldsymbol{k}(\tilde{x})\boldsymbol{C}_N^{-1}\boldsymbol{t}$$
$$\sigma_{\hat{y}}^2(\tilde{x}) = C(\tilde{x},\tilde{x}) - \boldsymbol{k}(\tilde{x})\boldsymbol{C}_N^{-1}\boldsymbol{k}(\tilde{x})$$

And $\boldsymbol{k}(\tilde{x}) = (C(x_1,\tilde{x}) \dots C(x_N,\tilde{x}))$ is the covariance between the training data and $\tilde{x}$ and $\boldsymbol{C}_N$ is the NxN covariance matrix of training data.

Problem:

Unlike a simple perceptron or even an LMS regression when one makes the transition from batch regression to online regression the general form of the regression does not change and if the same samples are chosen would result in the exact same results. However, when performing an online regression with GPR it is unclear what the best method is for choosing the next point. In our case we want to reconstruct an entire surface not just find a particular point so it would make sense that we should just find the next point with the largest variance. However, does this necessarily mean that if we choose this point we will make the best progress in the reconstruction? How do we define the "best" progress? Do some of the points we train with end up making our reconstruction worse? How do we deal with these?

Critical Review

This paper takes several approaches to the questions posed above.

Choosing the Next Point:

The paper suggestions two methods of choosing the next point. The first which the authors refer to as ALM (assumed to stand for active learning McKay) is the obvious one, compute the expected value and variance of a selection of points and choose the point which has the highest variance. Presumably this point will give the most information to the model or at least more than a randomly selected point.

The second method is slightly more complex but the general idea is the same. Here, instead of assuming the point with the maximum variance in the expected value will give us the most information, the expected change in variance is computed for each of the selected potential points. This method is referred to as ALC (assumed to stand for active learning Cohn). With the overall goal being to minimize the Mean Squared Error the mean square error is evaluated as decomposed into a variance and a bias term.

$$E_{MSE} = \sigma_{\hat{y}}^2 + E_x[(E_\tau[\hat{y}(x)] - y(x))^2]$$

If the model is correct or close to then the bias term becomes negligible and the variance term dominates. This makes the choosing future points possible because we no longer need to know their true value. As stated in the background, Gaussian process depends on the covariance matrix of all the points relative to each other, $K \; and \; K^{-1} * y$. The current covariance matrix is then appended and manipulated to represent the expected Covariance matrix for each of the potential points. This is below:

$$\mathbf{C}_{N+1} = \begin{bmatrix} \mathbf{C}_N & \boldsymbol{m} \\ \boldsymbol{m^T} & C(\tilde{x}, \tilde{x}) \end{bmatrix} \quad \mathbf{C}_{N+1}^{-1} = \begin{bmatrix} \left[\mathrm{C}_N^{-1} + \frac{1}{u} \boldsymbol{g}\boldsymbol{g^T} \right] & \boldsymbol{g} \\ \boldsymbol{g^T} & u \end{bmatrix}$$

$$\boldsymbol{m} = [C(x_1, \tilde{x}) \ldots \; C(x_N, \tilde{x})] \in \; \mathbb{R}^N$$

$$\boldsymbol{g} = \; -u\boldsymbol{C_N^{-1}}\boldsymbol{m}, \qquad \boldsymbol{u} = \left(C(x_N, \tilde{x}) - \boldsymbol{m^T}\boldsymbol{C_N^{-1}}\boldsymbol{m}\right)^{-1}$$

Finally, this is used to evaluate the expected change in variance for each of these potential points as below:

$$\Delta\sigma_{\hat{y}(\xi)}^2(\tilde{x}) = \; \sigma_{\hat{y}(\xi)}^2 - \; \sigma_{\hat{y}(\xi)}^2(\tilde{x}) = \frac{\left(\boldsymbol{k}_N\boldsymbol{C_N^{-1}}\boldsymbol{m} - C(\tilde{x}, \xi)\right)^2}{(C(\tilde{x}, \tilde{x}) - \boldsymbol{m^T}\boldsymbol{C_N^{-1}}\boldsymbol{m})}$$

The paper then goes on to briefly explain that this should be used in practice by averaging the variance over the selection of points and the point with the maximum change should be chosen as the next point to train on. As a side note the paper mentions that these methods of choosing the next point are not as interesting as they could be because they do not rely on any measured values, however this is in my opinion one of the most interesting parts and beauty of Gaussian process regression.

Evaluation of Choosing the Next Point Algorithms:

To validate and compare these results the authors tested the algorithms on two data sets, one which the authors created and new the underlying model and another from a simulation of the dynamics of a particular robot which they did not know the underlying model for. It is unclear exactly what they were predicting in the case of the robot. However the authors also leave out how they chose the random points to evaluate among for candidacy as the next point to train with. And nor do they suggest any methods or acknowledge that different methods may be used with different results. For example, in our physical case it makes sense for us to randomly sample points near the robots current location to evaluate their expected value to the regression and not the entire space for multiple reasons. Both methods appear to have potential for increasing the rate of the regression, however ALC appears to be superior on both data sets, especially the data from the robot arm. For us, ALC appears to be the natural choice unless its computational intensity slows down the program too much.

Bad Sample Points (Test Point Rejection)

The topic of this section is less clear but slightly surprising and more intriguing. The idea behind it is that some of the data points in your model may not be good for your generalization. Unfortunately the explanation is very poor but my understanding is that you can look back on the data you have and remove points which are causing or have a higher variance than others which have been evaluated. My reasoning for this is that even though the authors say "Given the true function and the GP regression" the authors follow this with "points from the test set are accepted or rejected according to the variance criterion." Additionally they perform test point rejection on the robot arm data set with which they do not have a true model. It appears that you look back on your model and remove some of the points that still have a high variance. The experimental results appear to work extremely well when the model is correct but not as well for datasets where the underlying model is unknown or incorrect. For both the toy data set and the robot dataset improvements are made most when the points are removed at low rates, for example if 10 appear to be bad for the model you would randomly choose 3 of these to remove. Additionally although not discussed, it appears from the figure on the robot data that as you reject more points your mean squared error will go down on average but that the variance will rise significantly after the first small level of rejection. This raises several questions, assuming their results are accurate and generalizable to other problems is there a better way to remove points other than simply by random selection from points which meet the criteria? The paper does not explore this or pose alternatives for methods of rejecting test points at all.

Thoughts and Impact on our project

At one point the paper discusses the value of being able to see how accurate the prediction is at a given point, and one could argue that the best improvement in the regression might be the one that causes the most points to have a very small variance rather than lowering the average variance or some other metric. The authors did not need to go into detail but I felt the paper would have been much more meaningful had they discussed why they made their choices and what other alternatives might have been for improving the regression. The paper could have gone into much more detail on its validation and explanations, particularly for the Test Point Rejection methods. Both methods appear to have merit for our problem particularly ALC although both will also need analysis for our project. Additionally, the idea of test point rejection is novel to us and further exploration through application in our project may prove fruitful. Lastly, although it is a given that we need to register our data for comparison to the CAD model of the surface and validation, exploring these options highlights the need for us to do this and not rely on our current visual comparisons.