Journal of Ryan Botts: Winter 2008

Ryan Botts

March 14, 2008

1 7 January 2008-16 January 2008: Cross Validation and Other Small Perfections

During this week I tried to wrap up a few things I have been working on over break. The code is working well, but now we need to perform some cross-validation. Cross-validation is a way to check that your algorithm works in a general setting and not only because you have "nice" data. We would perform a 10-fold cross-validation in which the data is divided into 10 groups. The training would be performed on 9 of the groups and tested on the remaining group. Repeat 10 times and you will know that most likely you have captured some underlying relationship in the data. Due to the fact that we only have 60 or so data points, we decided to perform a leave one out cross-validation. The disadvantage of having a smaller testing set simply that you do not have as many points to test on. In an algorithm such as this, you could expect that it would take a large number of data points to learn from, so you can not leave many out. The results came in for a cross-validation with a a rank 1 cube of size 3 separable function and there was a lot of variation in the trials. I am currently running the cross-validation with a rank 3 cube 3 separable function.

After looking back through the code and looking for ways to speed things up Dr. Mohlenkamp found that there was a nested loop in the ALS routine, which we had shown in the paper to be unnecessary. I rewrote this part of the code and found that speeds showed dramatic improvement. I am also running a test to check for fitting ability with the revised code. It seems that taking a little time away from the code has given me a chance to think about things differently and make many revisions I did not see before.

Beyond working with the code I have also worked on revising the paper. I feel like I am finally beginning to get a feel for mathematical writing and it is coming much easier. One technique which has been useful is getting a rough outline and gradually going back through the paper and modifying parts which need to be modified. I still need to add a few more statistics from the tests and make transitions a little smoother, but overall it is finally beginning to take shape.

In an effort to find several references I have begun reading several papers by Alex Zunger. I need to improve my technical reading abilities, but I am getting much faster at this. For next week I will continue reading about cluster expansions, and hopefully begin a little background work which will use similar techniques to fit data from the GRACE project.

2 17 January 2008-23 January 2008: Why Did the Cross-Validation Come out So Poor

This last week I edited the code to find the mean and variances for the cross-validation. I spent a lot of time on the paper. Edit. Edit again. Edit again. And repeat. It has been much more

difficult than I had expected to write a good math paper that is understandable to people without experience on the project.

I ran the cross-validation with smaller cube sizes and lower ranks and it appears that we were getting some amount of over-fitting due to the fact that when we lowered both of these the cross-validation resulted in smaller errors. I also read several papers by Alex Zunger about cluster expansions and genetic algorithms which I will use as references for the background in the paper.

This coming week I am going to work on getting better plots for the paper, running a few more tests where the separable functions are initialized to random values and hopefully get closer to a final draft of the paper. I am also going to try to write some code which will help tell us something about the most locations which are most important to the approximations.

3 24 January 2008-30 January 2008: How Many Revisions Can There Be?

This week has seen work resulting in very little visible progress. It seems as though the last final details are going to take a lot more work than you would expect. Every read through the paper finds 3 new paragraphs that need to be added and 3 that need major revisions. There is progress being made, though. The results of the cross-validation with very low rank and low cube size (2 and 2 respectively) were very good indicating that some amount of over-fitting was occurring at higher ranks and larger cube sizes. It also appears that in the cross-validation there are some essential structures which must be used in fitting the regression model.

I have spent a lot of time revising the tables so that they were dense in information and visually attractive. There is a lot more thought that goes into these than I expected. It is not nearly so easy as simply plotting your data points. It requires finding a good combination of dependent and independent variables which will show meaningful variations and possible trends in which structures are not fit well. For a regression problem such as this, it is not necessary to show that you have a really good fit, but showing poor fits may reveal reasons the model does not fit, allowing for future improvements. Plots of error may be used to find a correlation between goodness of fit and complexity of the structure or magnitude of energy. We have not found any of these as of yet, but we still might. A valuable lesson has been learned that in fact showing less good fits is in some cases more informative.

During this coming week I hope to be polishing the paper and figures. I will also begin the write-up for my presentation the following week. I also need to be finding a long-term project to work on once this one is finished.

4 1 February 2008-6 February 2008: More Revisions and a Poster Presentation

This week has been an editing week. More revisions have been made to the paper, and it is getting better although there is still much work to be done. The explanations are better, but it needs to be more cohesive and smooth. This is probably the result of working on individual pieces one at a time. I am still looking into good ways to produce graphics of these structures, but need more time on this.

I also began a poster presentation of this research which is a much different beast than writing a paper. It must be very concise and for this one, to a broad audience. These factors have posed many problems to the write-up, but I think they have been overcome.

Hopefully the result of this next week will be a finalized version of the paper and the final bits of cross-validation of the code.

5 7 February 2008-12 February 2008: What you can learn from a presentation

This week I spent most of my time preparing for a seminar presentation. I had always viewed presentations as your chance to share what you have done and hopefully provide tools other people could use. My views on this have changed.

First, the presentation requires that you draw the essential aspects from your project and connect them in a very logical manner. We forget that once we work with something for a while we no longer need to work on the problem sequentially, but may freely jump from place to place sometimes forgetting why we did things in a certain way. Presenting the material requires that you go from the beginning through the end in its entirety with no jumps. It forced me to rethink why we did many things the way that we did.

Although frequently dreaded, presentations such as this are valuable in the acquisition of new viewpoints. They give the opportunity to gain expert opinions on things that you missed. For example I had not thought about how much stronger and concrete I could make my conclusions by performing a more thorough statistical analysis. I had never thought about things like confidence intervals, etc. The outside perspectives gave me many ideas for new ways to analyze our results.

During the coming week I would like to speak with a statistician to learn techniques which would be best suited for analyzing this type of data. These tools would show much more concretely the success of our machine learning procedure. They would also be very useful for future projects.

I also spent time this week on other things such as learning new tools in LaTeX and obtaining more cross-validation results. I reduced the amount of randomness in the initialization for the cross-validation and obtained very accurate results.

6 13 February 2008-20 February 2008: Second Opinions

This week I met with Dr.Wei Lin, a statistician from our department to discuss how to best report our results as the statisticians seemed to have the most questions about the presentation. The main concern seems to be that we obtain a fit that is too "good," especially when considering the number of data points we have to train from. We may explain the quality of our fit in two ways: the first is that we do not actually have 57 data points, but we have $\sum_{x \in D} |H_x|$ points we may use for training, and secondly it is quite likely that the unknown function we are fitting is very simple. In order to verify that our data is "nice" enough he suggested that we add some known amount of random variance to our data so that when fitting we now attempt to fit something of the form $f(x_j) \approx y_j + \epsilon$. If our model is fitting as it seems to we should still approximate the y_j, however we will not be able to fit the random variations ϵ. As we know how much random error we have we should observe that our model now fits with larger errors and we should be able to predict how much error that will be.

I ran the code with too large of values for ϵ first and it could not fit the data, so we at least know it will break. I will run it again with smaller randomization. I am doing a little more analysis of the results from the initial training data and then we should have all of the information we need for the paper.

I have begun to look into gene regulation networks which may be modeled as coupled system of ODE's. These appear to be something we could model using regression in higher dimensions. I am

attempting to find data sets for this and to learn more about what has already been done. I am also looking into a better way to say that the consistency operator does actually enforce consistency.

7 21 February 2008-27 February 2008: More Revisions

This week worked more on editing the paper. The idea of consistency has been difficult to convey clearly. We have also had to make many changes to the way we describe our results. Currently we are going to do a little more work on the consistency and then we need to figure out why the mean of the max errors in the cross-validation appear to be better than the errors in the initial testing. The error in the basic test is within one standard deviation of the error in the cross-validation hence it is plausible, but if I run it again I might see that the basic test comes out better. I have checked that the multiple cross-validation results come out with similar values and they are all very close. I am going to try running the basic test a few more times to see if anything can be found.

I have also begun to look into gene regulatory networks. There are two models for GRN's which might be of interest to us. The first is using a coupled system of ODE's. If X_i is the concentration of protein or mRNA i, then we may consider the effect of the concentrations of all different substances on 1 of these using the equation: $\frac{\partial X_i}{\partial t} = f_i(X_1, X_2, ..., X_n)$. The f_i are unknown functions of many variables and would hence lend themselves to a regression procedure such as we have used for the crystal structures. The second representation of these is using Boolean functions, this also may be of some interest to us as these are unknown functions of many variables from the set of values $[0, 1]$. The presence or lack of proteins or mRNA regulates the expression of other genes, and hence the presence of other proteins or mRNA. If we let $\mathbf{x} = (x_1, x_2, ..., x_K)$ be a vector representing the presence of protein x_i, then we may model the network using $\frac{\partial f(\mathbf{x})}{\partial x_j} = f(\mathbf{x}^{(j,0)}) \oplus f(\mathbf{x}^{(j,1)})$, where the vector $\mathbf{x}^{(j,i)}$ represents the vector $\mathbf{x}$ with the two possible values 0 or 1 at entry x_j. So the $\frac{\partial f}{\partial x_j}$ represents the change in f caused by the change of states to x_j. I can not find much about f, although it is somehow related to the state of the system. I did however learn what is meant by the derivative of a Boolean function. If we consider a function $f : \{0, 1\}^N \to \{0, 1\}$, then the partial derivative $\frac{\partial f(\mathbf{x})}{\partial x_j} \in \{0, 1\}$ and represents whether changing the state of component x_j changes the state of f.

Some time today Dr. Just thought he would be able to get me some data, or some benchmark networks used in the differential equation approach. I hope that we will be able to modify our code to learn what these networks are.

8 28 February 2008-4 March 2008: Finalizing the paper and GRN's

During this past week we finished editing the paper and sent it off to get feedback on the results. We fixed the problems between the cross validation results and the fitting results. If the results are not sufficient in practice I learned about a method which would overcome that problem and yet not over fit: regularization. You may increase the cube size, but you penalize for using more terms than necessary. Dr. Mohlenkamp had explained this to me in the past, but I didn't understand it until I read his classification paper.

I finally received some synthetic GRN data to work with. During the coming week I am going to begin working with the GRN data. I will also be cleaning up all of our code in preparation for passing it on to the physicists.

9 5 March - 11 March 2008: GRN data processing and a few thoughts on Clusters

This week was less productive than others due to the large end of the quarter to-do list, which includes the final report, however, I made progress on a few things. I have developed the code to process a list of files containing concentrations used in gene regulation networks. The next step here will be to apply the rest of the machine learning procedure to this data. This data does not have any symmetries as the crystal's did, so this procedure should be relatively straight forward.

In addition to this, I performed a few more tests on the crystal structures. We performed the cross-validation with a rank 2, cube 3 function and checked that rank 2 cube 5 over-fit. We have never justified the use of these parameters, and do not even know that they are optimal. To justify this, I went ahead and performed a cross-validation for all cube sizes and all ranks. We should then be able to see where the errors over the testing data are approximately equal to the errors over the training data.

I also modified the code to use VectorDiff so that we can use regularization to force the algorithm to use the least number of terms possible. I do not know what good values should be for the regularization parameter so I began runs for several different values. We should see that now we can obtain good cross-validation results for larger rank and cube size. By next week I should be able to compare these with the original results.

Finally, during this past week, I was able to plot the sites and the magnitude of the interaction at each site. I did not learn anything new from these plots, but with the regularization we might be able to force our function to only use the most essential sites in the computation which could possibly allow us to locate sites involved in important interactions.