**CECS 551 Final Project Diary**

**04/24/2017**

1. I started narrowing down possible project paths. I knew that I wanted to explore topics in unsupervised learning, but it was difficult deciding what to do.
2. My research is in the field of bioinformatics. I am given the task of finding methods to cluster large, complex and rich biological data sets.
3. Thus, I decided to explore some theoretical foundations of unsupervised learning. Although I do not have to explain the theoretical depths of unsupervised learning in my thesis and potential publications, this project would be an incredible opportunity to explore some of the underlying mechanisms of such methods.
4. I decided to explore the following topics: latent variables, expectation maximization and mixture of gaussians, and autoencoders.
5. I also decided to apply k-means clustering to a protein data set in a distributed Hadoop cluster that I have built.

**05/01/17**

1. Although I had already built a Hadoop cluster, I decided to scratch everything and start over. This is mainly because there was a major update that fixed many bugs. I also wanted to add a few more slave nodes.
2. The main complication with my desired set up was that I wanted my main lab computer to be the master node. However, my lab computers are in their own private network with limited root access.
3. Nonetheless I fresh installed Debian on all of the lab computers and connected all of them via a network switch.
4. I downloaded and configured Hadoop on the master node.
5. I then set up passwordless SSH to all the slave nodes from the master node.
6. There were many networking complications that I ran into. These are not worth mentioning, but they were problems I had never encountered.
7. I was not able to finish the set up today. It was 1am and I was no longer able to stay awake.

**05/02/17**

1. I came right back into the lab the next day with very little sleep. I had to get this thing done today.
2. I learned that simple network switches do not conveniently assign IP addresses to computers. I had to set up my own DHCP server on the master node.
3. After hours of research I learned that I need a router. Routers have DHCP servers built into them. I found an old, unused router and eventually got the network set up.
4. I could now launch Hadoop and all the respective slave daemons from the master node. I was also able to monitor the status of slaves using Hadoop’s built-in web UI. This made me very happy.

**05/07/2017**

1. With the Hadoop cluster up and running, I decided to finally start on the write-up of my project.
2. I gathered my initial set of sources:
   1. **Machine Learning by Peter Flach**
   2. **Deep Learning by Ian Goodfellow**
3. I have decided to migrate over to Spark (as opposed to Map Reduce).
   1. Datasets are smaller
   2. Data processing/analysis is done in memory
   3. Spark runs up to 100x faster than MapReduce in memory or 10x faster in disk
   4. Provides high-level APIs in Java, Scala, Python, and R
   5. Has a well-known machine learning library called MLlib
4. I should mention that Spark still runs in a Hadoop cluster.
5. I downloaded, installed, and configured Spark.
6. By the end of the day I was able to successfully run Spark in the Hadoop cluster.

**05/08/2017**

1. I established the sections of my final report.
2. I started and completed introduction section of my report.
3. I gathered a few more resources:
   1. **Deep Learning; Nature; LeCun**
   2. **Latent variable modeling; Shangai Archives of Psychiatry; Li CAI**
   3. **Overview of K-means and Expectation Maximization Algorithm for Document Clustering; Umale**
   4. **Clustering With EM and K-Means; Alldrin**
   5. **Other class notes from credible university courses (UCSD, CMU, Stanford)**
4. I started and completed the latent variables sections of my report.

**05/09/2017**

1. I started taking a deep look at expectation maximization.
2. I read the EM section in Machine Learning by Flach.
3. I started working on the derivation for the EM algorithm.
4. I updated my report with my newly acquired knowledge.

**05/10/2017**

1. I completed the derivation of the EM algorithm.
2. I started the application of EM to a Gaussian mixture model.
3. This derivation was quite complication so I was only able to finish about half of it.
4. I updated my report with my newly acquired knowledge.

**05/11/2017**

1. I completed the derivation for the application of EM to the Gaussian mixture model.
2. I also showed that k-means is a special case of EM.
3. I then updated my report.

**05/14/2017**

1. I ran k-means locally 30 times, each time with a different number of starting clusters.
2. I ran k-means on Spark 30 times, each time with a different number of starting clusters.
3. I created plots of the running times vs. number of clusters.
4. The results are quite impressive; I discovered that Spark is an order of magnitude fast.

**05/15/2017**

1. I put together my presentation.
2. I realized that it would be beneficial to briefly go over Hadoop and Spark as well as my journey into building a Hadoop cluster.
3. The presentation now only needs screenshots of the Web UIs for Hadoop/Spark as proof of a running cluster.
4. I made the final edits to my report and put everything together.