Using Byzantine Statistics to Approximate Future Climate

1. **Introduction**

The basis behind my project was that climate is predictive of itself. When you look a long-term average graph of various months over the course of several years it is predictable and the natural fluctuations that can be seen in these graphs are predicable to a certain extent based on the previous years and months of climate. It was my hope that applying byzantine statistics in the form of a deep neural network, I would be able to predict future climate trends up to a year in advance.

1. **Methodology**
   1. **Data Collection and Dataset Construction**

Due to the general lack of accurate climate data going back more than about 120 years, I used several archived weather stations from across Canada. Generally, I found that airports had both the most consistent and longest lasting historical climate data. When it came to the construction of my datasets, I tried several different structures in terms of their architecture and general data makeup. I tried using data from all across Canada, attempting to use as even of a spread of latitudes as I could in the interest of making my network as general as possible. In addition to this I also used a dataset take exclusively from the prairies to see if the network would perform better if the data came from a consistent climate. In terms of the architecture of the individual lines of the dataset I used one with 5 years of monthly input data and one with 10 years’ worth of input data. Both of these architectures had 1 year worth of output months. It is important to note that I was using both temperature and precipitation data for all months of data I used. I also tested a variation on the datasets where I used adversarial data training by taking a full dataset, duplicating it, and slightly modifying all the values to see if that made my prediction calculations more robust.

* 1. **Network Architecture**

My network had an input shape equal to 24 times the number of input years, three hidden layers in descending order with sizes of 120, 60, and 30, and due to a limitation of Keras that I was unable to overcome, it had 1 output node and I used a different network for each target. I chose this design to do my best to reduce the possibility, or at least extent, of overfitting while still allowing for the large input shapes.

1. **Experiments**
   1. **Canada**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Month | January | | February | | March | | April | | May | | June | | |
|  | Tracked | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre |
|  | SD | 6.94 | 79.26 | 6.32 | 63.97 | 5.00 | 56.56 | 3.20 | 49.31 | 2.97 | 41.61 | 2.93 | 42.30 |
| Normal | 5 in | 3.01 | 35.22 | 2.98 | 32.67 | 2.12 | 25.28 | 1.69 | 28.45 | 1.54 | 37.31 | 1.85 | 38.79 |
| 10 in | 2.98 | 27.83 | 2.59 | 28.85 | 1.99 | 24.05 | 1.48 | 26.41 | 1.52 | 33.66 | 1.82 | 39.58 |
| Augmented | 5 in | 1.97 | 16.74 | 1.97 | 18.06 | 1.58 | 14.10 | 1.18 | 13.99 | 1.24 | 18.92 | 1.40 | 22.99 |
| 10 in | 1.88 | 15.87 | 1.74 | 13.22 | 1.40 | 11.41 | 1.18 | 14.83 | 1.38 | 18.10 | 1.44 | 21.64 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Month | July | | August | | September | | October | | November | | December | | |
|  | Tracked | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre |
|  | SD | 2.50 | 40.72 | 2.38 | 46.46 | 2.66 | 48.26 | 2.78 | 69.51 | 4.59 | 82.16 | 6.11 | 82.98 |
| Normal | 5 in | 1.98 | 31.53 | 1.92 | 39.01 | 1.80 | 33.68 | 1.59 | 38.82 | 1.88 | 39.59 | 2.30 | 35.21 |
| 10 in | 1.92 | 33.68 | 1.76 | 38.71 | 1.60 | 31.11 | 1.48 | 27.44 | 1.76 | 34.43 | 2.41 | 25.58 |
| Augmented | 5 in | 1.48 | 19.62 | 1.62 | 21.31 | 1.43 | 20.08 | 1.27 | 20.95 | 1.39 | 20.30 | 1.53 | 19.05 |
| 10 in | 1.48 | 17.39 | 1.39 | 19.43 | 1.31 | 17.26 | 1.17 | 15.06 | 1.50 | 17.34 | 1.52 | 18.17 |

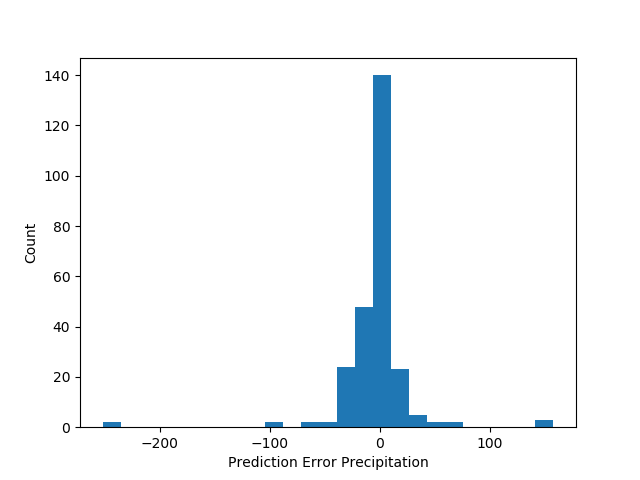
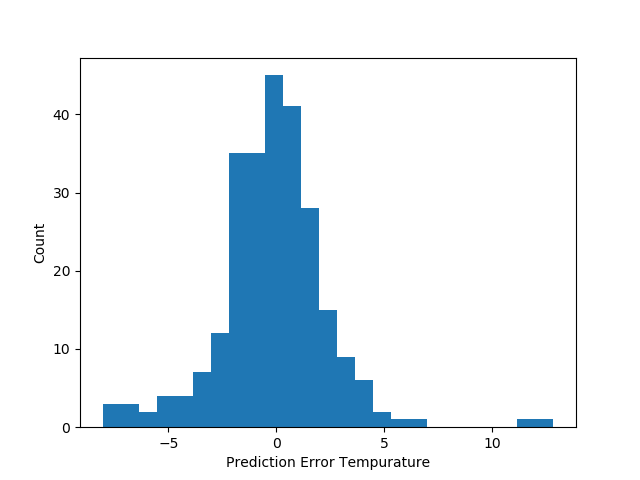
* 1. **Prairies**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Month | January | | February | | March | | April | | May | | June | | |
|  | Tracked | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre |
|  | SD | 6.37 | 13.93 | 5.61 | 10.20 | 4.33 | 12.05 | 2.86 | 18.73 | 2.08 | 32.42 | 2.02 | 43.63 |
| Normal | 5 in | 2.74 | 10.49 | 2.62 | 7.90 | 2.05 | 10.16 | 1.36 | 14.61 | 1.49 | 27.70 | 1.59 | 38.53 |
| 10 in | 3.09 | 8.91 | 2.62 | 7.72 | 1.70 | 9.49 | 1.29 | 13.82 | 1.19 | 23.49 | 1.88 | 40.24 |
| Augmented | 5 in | 1.84 | 5.03 | 1.57 | 4.46 | 1.38 | 5.27 | 1.30 | 7.52 | 1.24 | 13.12 | 1.31 | 17.19 |
| 10 in | 1.53 | 4.79 | 1.34 | 4.46 | 1.32 | 5.78 | 1.09 | 7.21 | 1.17 | 10.99 | 1.30 | 18.97 |

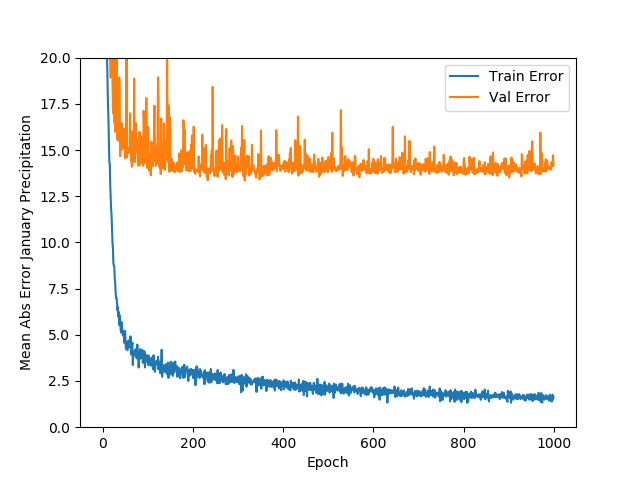
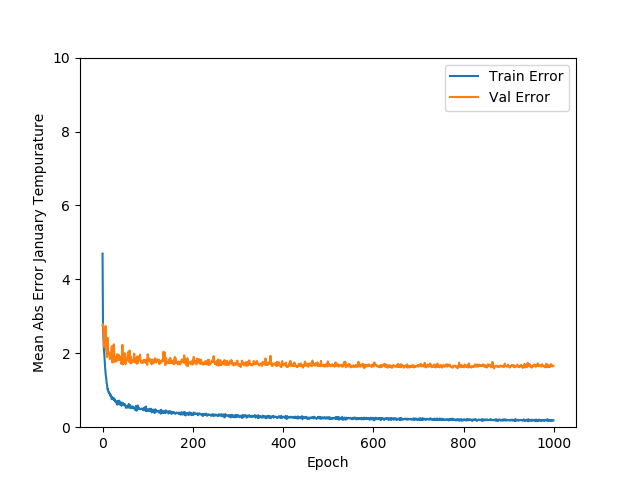
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Month | July | | August | | September | | October | | November | | December | | |
|  | Tracked | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre | Temp | Pre |
|  | SD | 1.93 | 39.33 | 2.26 | 38.97 | 2.24 | 28.76 | 2.46 | 18.28 | 4.34 | 14.35 | 5.58 | 11.89 |
| Normal | 5 in | 1.80 | 35.29 | 1.68 | 27.68 | 1.45 | 22.21 | 1.34 | 12.84 | 1.99 | 11.00 | 2.70 | 8.81 |
| 10 in | 2.00 | 33.76 | 2.01 | 26.99 | 1.51 | 21.11 | 1.32 | 13.91 | 1.70 | 10.10 | 2.54 | 7.62 |
| Augmented | 5 in | 1.44 | 15.52 | 1.38 | 15.24 | 1.26 | 11.33 | 1.12 | 6.57 | 1.33 | 6.06 | 1.78 | 4.75 |
| 10 in | 1.38 | 15.25 | 1.31 | 14.30 | 1.27 | 10.89 | 1.11 | 6.61 | 1.28 | 5.34 | 1.64 | 5.05 |

1. **Analysis**

What these experiments prove is that climate, both local and in a wider area are very predictable via byzantine statistics. That being said, there are a few problems with my network and analysis. The first one is that my network does not deal with outliers very well.



As you can see, most of the predicted values are quite accurate, but there are some extremely inaccurate values. Unfortunately due to the nature of byzantine statistics, outliers are very difficult to predict, and in the case of climate where factors that might not be initially noticeable in a local area can have a massive effect very quickly, my network does not deal with it very well. In addition to this, my network is prone to overfitting.



The best way to reduce things like overfitting is to get a better data set. Unfortunately this is really an option for me. Due to the nature of long term local weather and human record keeping, there is little if any data before about 150 years ago, and the older the data is the more inconsistent it is. I could build a network that allows for incomplete input datasets, however I ran out of time to attempt something like that.

1. **Conclusion**

Ultimately I’d say that this was a successful project. I learned much about deep learning and its applications. I’ve also proved an unexpectedly strong correlation in terms of climate. If I had the opportunity to take this course again I would like to figure out how to make multiple outputs work and also find a way to get a better data set. Having said all of this, I would like to thank you for being my supervisor for this independent study. It was everything I could have hoped for and more, and it was a fantastic way to round off my undergraduate degree.