CSCI 5622: Final Project Proposal: Predicting Rainfall in California with Neural Networks

Minah Yang, Kris Tucker,
Batzolboo Batsaikhan, and Will Shand

February 2018

1 Introduction

Droughts in California are frequent and have profound economic and environmental impacts. For example, the recent drought from 2011 to 2017 led to more than 100 million trees dying, water restriction orders within California, and a costly blow to the 44.7 billion dollar agricultural industry that produces half of the vegetables, nuts, and fruits in the U.S.[2] Despite the copious amounts of data gathered from the many droughts in recent history, predicting rainfall is still a difficult problem. The highly chaotic and variable nature of the weather makes any kind of prediction difficult, especially over longer periods of time.

Our goal with this project is to gain a better understanding of how modern machine learning techniques can be adapted to assist meteorologists. In particular, we aim to look at neural nets that are able to look at spatially distributed time series data with highly complex features to determine its accuracy at predicting rainfall in California.

2 Possible Approaches

Neural networks are one of the current state-of-the-art machine learning tool for processing data with many highly complicated features. We briefly cover three broad classes of networks that could be useful in achieving our goal.

2.1 Deep Neural Nets

In class we have discussed classifiers with linear decision boundaries, such as logistic regression or support vector machines. These types of classifiers have an obvious limitation, however, in that they cannot handle data that is not well separated by hyperplanes. The complexity of our data is such that a simple linear classifier is likely to be insufficient, providing motivation for the use of *deep neural networks* (DNNs).

DNNs can be difficult to train, however, and because of the fully connected nature of each pair of layers, the number of weights may grow to an extent that the traditional training process involving back propagation proves intractable. Our data set includes several features that take the form of a grid over large regions of the globe, for example ocean surface temperature and pressure. This makes for a very large feature set, with correlations that may be limited in geographic extent. To account for this, we are considering using either convolutional or recurrent neural networks, which are the topics of the next sections.

2.2 Convolutional Neural Nets

Convolutional neural networks (CNNs) have been used with great success in analyzing data that are distributed across a grid, most notably in the field of computer vision. Part of our project will be to evaluate the viability of CNNs in meteorology where their applicability is still largely unexplored.

2.3 Recurrent Neural Nets

Classic feedforward networks can be visualized as directed acyclic graphs with weighted edges. On the other hand, models in the class of *recurrent neural networks* (RNNs) allow neurons to pass inputs to previous neurons in the network, relaxing the requirement that the network be acyclic. Furthermore, they allow a dynamic internal state, which makes them excellent for processing time series data.

3 Logistical Details

The data used for this project concerns California rainfall and comes from the 7th International Workshop on Climate Informatics last year. We were made aware of this data and given access to it by a contact at NCAR.

To perform our analysis we will evaluate the performance of different neural network models on these data using the TensorFlow framework [1].

We hope to be able to compare our results with the results from the workshop that was held last year, which had focused on coupling dimension reduction and logistic regressions.

References

[1] Abadi et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *Google Research*, 2015.

[2] California Department of Agriculture. California agricultural production statistics: 2016 crop year report, 2017.