Weather Forecasting using Probabilistic Graphical Models

Felipe Hernandez FELIPEH@ANDREW.CMU.EDU
Machine Learning Department, Gates Hillman Center

Amos Ng AJNG@ANDREW.CMU.EDU
Language Technologies Institute, Gates Hillman Center

Abstract

Weather prediction has usually involved running physical models of weather phenomena in order to predict future conditions. In this project, instead of focusing on the physics, we propose using probabilistic models based on meteorological statistics gathered by NASA to produce future weather conditions.

1. Introduction

Weather forecasting is used for a broad range of purposes, ranging from personal activity planning to large-scale economic decision-making to emergency preparation and response. The availability and accuracy of forecasts thus have a profound impact on human activities at many levels, both in measurable and unmeasurable aspects.

However, predicting weather is a difficult research problem. Most often, physically-based models with global and regional scales are used to forecast future conditions. In this project, we will instead take a probabilistic machine learning approach focused on a regional scale and predict atmospheric variables at specific geographic locations.

The forecasts will be based on prior atmospheric states in the neighborhood of the selected location. In particular, we will attempt to predict the probability distributions of variables such as pressure, precipitation, and temperature based on data recorded by NASA using assimilated land products.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

2. Related work

Researchers in the atmospheric sciences have investigated a variety of methods for weather forecasting. Many subtle physical phenomena affect the weather at any given time, including energy and mass transfer between the sun and different layers of the atmosphere, ground, and ocean. Machine learning approaches have been explored to construct simplified models based on atmospheric measurements, but are not very popular among meteorologists.

Many of these techniques are tailored to fit the nature of the observations available. Weather monitoring stations are the predominant data source, providing point measurements with high accuracy and varying temporal resolution. Artificial neural networks have proven to be effective in such cases for forecasting rainfall amounts in the near future given a time series of previously observed values (Maier & Dandy, 2000). More recent works have attempted to combine NNs with other techniques to improve the performance of predictions. In (Hong, 2008), a recursive NN is trained using a support vector regression together with a chaotic particle swarm optimizer. Other algorithms used in weather forecasting include linear regression, discriminant analysis, logistic regression (Applequist et al., 2002).

Recently, meteorological observations from satellite and Doppler radar have become widely available, providing ubiquitous coverage at the cost of decreased precision. These sources of information add spatial dimensions to the forecasting problem. For example, Fourier spectrum, structure function, and moment-scale analyses are used to understand radar precipitation in (Harris et al., 2007). Forecasting models have also been created to take advantage of interdependencies between atmospheric variables that are measured or estimated using other models. Rain-gauge data and outputs from atmospheric models are used for forecasting precipitation in (Kuligowski & Barros, 1998) and

(Ramirez et al., 2005).

3. Data sets

As mentioned previously, there are multiple sources of meteorological information available online. Usually these sources are provided by government agencies such as NOAA and NASA, and have different levels of post-processing of raw data from gauges, land radars, and satellites.

The North American Land Data Assimilation System (NLDAS), a service hosted by the Goddard Earth Sciences Data and Information Services Center at NASA, is a multi-variable source of information produced through the assimilation of land measurements. We will focus mainly on this dataset due to both its relative high precision and the multiple variables it reports: precipitation, atmospheric pressure, humidity, temperature, wind speed, convective potential energy, and radiation flux.

The NLDAS product provides hourly weather data for the US beginning in 1980. Each sampled time in the data set includes the following values per cell in a grid of resolution 1/8 of a degree in both latitude and longitude directions. Figure 1 provides an illustration of one such hour-long sample.

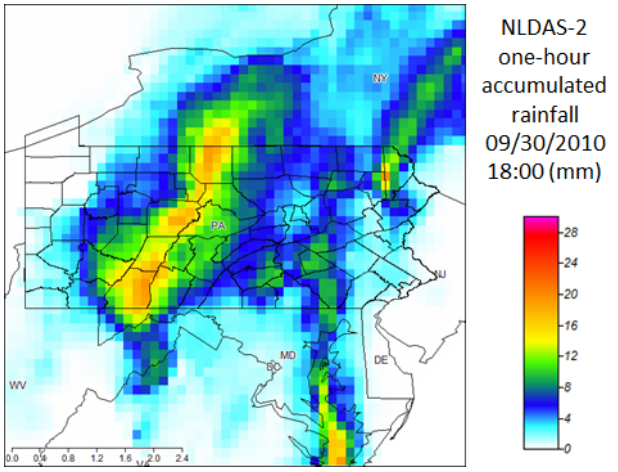

Figure 1. This is one example of a rainfall image produced during a severe storm in Pennsylvania in 2010. Each cell in this image has a size of roughly 10 km × 10 km. More information on the NLDAS-2 data products can be found on the following URLs: `http://ldas.gsfc.nasa.gov/index.php` and `http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings`.

4. Activity plan

We plan to use different probabilistic graphical models to try to determine the probability distributions of weather variables maps given the observed values from previous time steps. We will first use variable discretization so that discrete models can be used. Afterwards, we will attempt using models that are able to handle continuous variables.

To establish a baseline approach, we will first use simple graph representations, such as Nave Bayes networks. Later, we will attempt to use models with more complex topologies that take into account the interdependencies of variables.

Some of the activities that we will perform include

1. Selection of the dataset and the study area
2. Data download and pre-processing
3. Variable discretization strategy
4. Naïve Bayes formulation
5. Review of probabilistic graphical models for continuous variables

References

Applequist, S., Gahrs, G. E., Pfeffer, R. L., and Niu, X.-F. Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather and Forecast*, 17(4):783–799, 2002.

Harris, D., Foufoula-Georgiou, F., Droegemeier, K. K., and Levit, J. J. Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, 2(4):406–418, 2007.

Hong, W. C. Rainfall forecasting by technological machine learning methods. *Applied Mathematics and Computation*, 200(1):41–57, 2008.

Kuligowski, R. J. and Barros, A. P. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and Forecasting*, 13(4):1194–1204, 1998.

Maier, H. R. and Dandy, G. C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15:101–124, 2000.

Nasseri, M., Asghari, K., and Abedini, M. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert Systems with Applications*, 34:1415–1421, 2008.

Ramirez, M. C. V., de Campos Velho, H. F., and Ferreira, N. J. Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. *Journal of Hydrology*, 301:146–162, 2005.