When to Picnic?

Peter Barnum and Vinithra Varadharajan
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213
pbarnum@cs.cmu.edu,vvaradha@ri.cmu.edu

1 Introduction

The quantitative and reliable prediction of the level of precipitation is important for scientific, economic and ecological reasons [3]. Currently, there are many global climate models, but there is still a great deal of uncertainty in their predictions. Not only do they not agree with themselves, they do not even predict the past correctly [7]. (This is a similar idea to having poor training error.) Such poor performance rates are primarily due to the vast number of local and global factors that influence weather. Hence, the need to accurately predict precipitation levels given the datasets of historical records from different areas is an ideal machine learning problem. Many supervised learning techniques are applicable, with different levels of accuracy. In addition, the choice of a technique also needs to take into consideration that these datasets often suffer from missing values.

This report discusses the use of one such machine learning technique, namely nearest neighbor, to predict the level of precipitation expected on a particular day at a particular location, given data on the level of precipitation that occurred on previous days at the same location and at neighboring locations.

We begin by stating the problem and describing the data provided. Discussing previous work on the topic sets the scene for our approach to the problem by discussing previous work on the topic. This is followed by a description of the nearest neighbor machine learning technique and how it is used in weather prediction. We then describes the design and implementation of experiments and discuss the results obtained. The report ends with a discussion of future work and conclusions.

2 Problem definition

We use a machine learning technique to predict the level of precipitation based on historical precipitation data. We use the Widmann and Bretherton dataset that includes 45 years of daily precipitation data across 50 km x 50 km from the Northwest of the US in netCDF format. The data has three dimensions: latitude, longitude and time in days. The unit for each entry is mm/day and refers to the precipitation that occurred at a particular location, specified by the latitude and longitude values, and on a particular day, specified by the time value. Such an objective requires understanding the data and its features, selection of a machine learning technique, application of the technique by making assumptions and evaluation of the entire approach by analyzing the results.

The data extracted from the netCDF file is scaled by a factor of 0.1 and has several missing values. The missing value is indicated by a value of 32767. The data has been prepared by descaling the data and setting the missing value entry as 0 instead of 32767. Dealing with missing data explicitly adds unnecessary complexity. Given this, we have no intelligent way to pick a prior, except that there is more often no rain than some rain.

3 Related Work

People have tried to predict weather with various techniques and models for millenia. Early weather prediction algorithms involved memorizing lists of predictive algorithms, such as "red sky at night, sailor's delight", which was probably based on a data driven approach that recognized that if the sky was red the night before, it often rained the next day. Advances in modeling have led to additional techniques. According to Beniston [1], a variety of models are used, based on the resolution that is needed. For example, much more specific physical effects are used for local weather prediction, compared to global climate prediction. Wikipedia [12] divides these precisions into two categories, Global models and Regional models. Common global models are GFS, NOGAPS, GEM, ECMWF, UKMET, and GME. Common local models are WRF, NAM, NMM-WRF, AR-WRF, MM, and HIRLAM. These models predict a variety of factors, such as temperature, dew point, wind speed and direction, precipitation, and precipitation type. In contrast, our work is only trying to predict the amount of precipitation. We have at our disposal only a smaller set of features than those that these models take advantage, so we cannot use these models directly.

Many different machine learning methods and assumptions have been suggested to predict weather and their accompanying difficulties have been listed. In [4], Palmer approaches the problem of uncertainty in forecasts of weather and climate using ensembles of integrations of comprehensive weather and climate prediction models, with explicit perturbations to both initial conditions and model formulation resulting in an ensemble of forecasts that can be interpreted as a probabilistic prediction. He then uses singular-vector methods to determine the linearly-unstable component of the initial probability density function. He bases his prediction systems on timescales of days, seasons and decades. He states that many of the difficulties in forecasting predictability arise from the large dimensionality of the climate system. In [7], the Bayesian approach to model-based data interpretation has been used to investigate global climate modeling and prediction. It has been found to be particularly useful in applications where a large amount of prior domain knowledge is available. The Bayesian approach can not only find the most probably model, but it can also say how accurate the prediction is. Two methods that have been suggested specifically related to prediction of precipitation are neural networks [5, 9, 2] and prognostic equations [10]. In [3] Ehrendorfer states that quantification of atmospheric predictability asks for the rate at which two initially close trajectories diverge for given atmospheric dynamics. Such estimates place upper bounds on time horizons over which useful forecasts may be expected. The literature stated here has led us to believe that given our dataset two key features worth analyzing are the influence of time and space on the precipitation at a particular location.

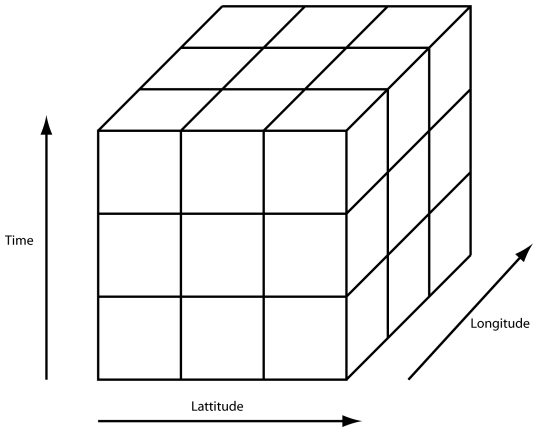
4 Using nearest neighbors for prediction

As discussed above, weather prediction is a complex and unsolved problem. The problem is complex, largely due to the huge number of hidden factors. It would not be unreasonable to say that if weather was a graphical model, there would be a million latent variables for every observed one. Given this complexity, we do not want to try to use a parametric method, as it is not clear that such a model would be able to capture subtle underlying causes that lead to the weather. Using more data-driven approaches versus physics-based modeling has been a trend in recent years, due to the complexity of the models, the chaotic nature of many physical phenomena, and availability of powerful computers [8]. We use a non-parametric technique, nearest neighbor, to predict the weather in a given day, given knowledge of the precipitation for some number of previous days. The advantage of data-driven methods, such as nearest neighbor, is that they allow for arbitrarily complex patterns to be captured, given sufficient data. This is doubly true, as we are attempting to predict based on just the amount of precipitation per day over a period of years, rather than explicit measurements of contributing factors, such as air pressure and terrain. Because we do not have information for many of the contributing factors, it is not possible to construct a model to take advantage of them.

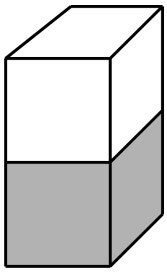
We use two different methods for predicting rainfall, unweighted nearest neighbor with space-time blocks, and weighted nearest neighbor such that locations that are further away in space-time are given less weight than those close by. Both methods predict the weather of a day in the same manner:

1. Identify the day (A) for which the precipitation needs to be predicted.

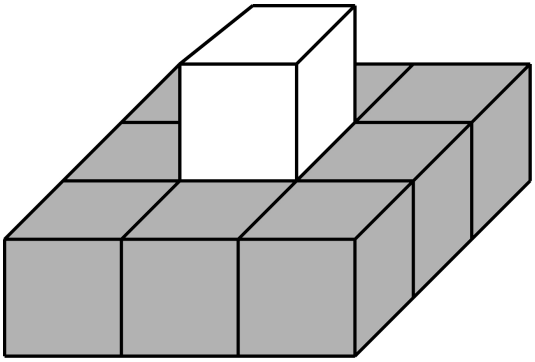
2. Identify a set of space-time blocks (B) associated with that day.
3. Find the set of space-time blocks (C) from the dataset that is the closest match to the identified set of space-time blocks (B).
4. Predict the precipitation of (A) to be the same as that of the corresponding day (D) associated with the closest matching block (C).



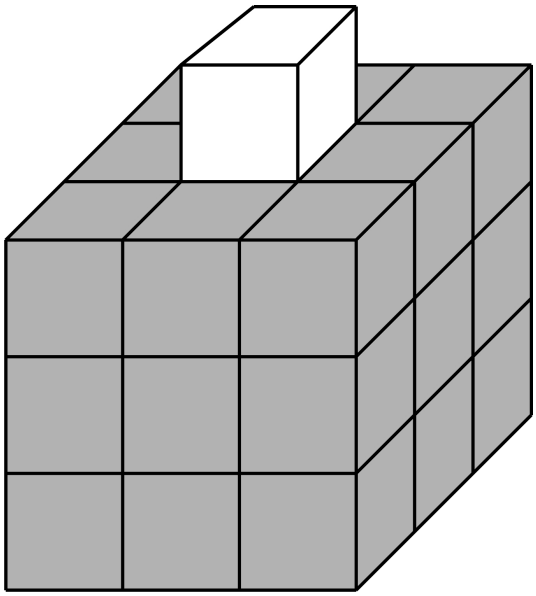
(a) A block in space and time. The width and depth of the block represent spatial locations and the height of the block represents different points in time, ranging from older time periods at the bottom to later time periods at the top. The top most layer represents the present day .



(b) Inferring the queried (white) value based on it's own value on the previous day.



(c) Inferring the queried value based on values at several neighboring locations on the previous day.



(d) Inferring the queried value based on a block of previous days and locations.

Figure 1: Methods of inference

4.1 Traditional nearest neighbor

Unweighted nearest neighbor with space time blocks is illustrated in Figure 1. The most basic case is simply trying to predict the rainfall in a certain location, with only the information of what happened the day before at the same location. With reference to the algorithm presented in the

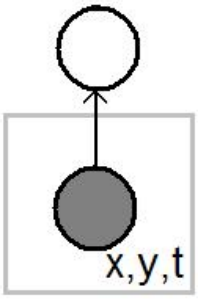


Figure 2: The graphical model for inference with a space time block. In this case, latitude is denoted by x, longitude by y, and time by t.

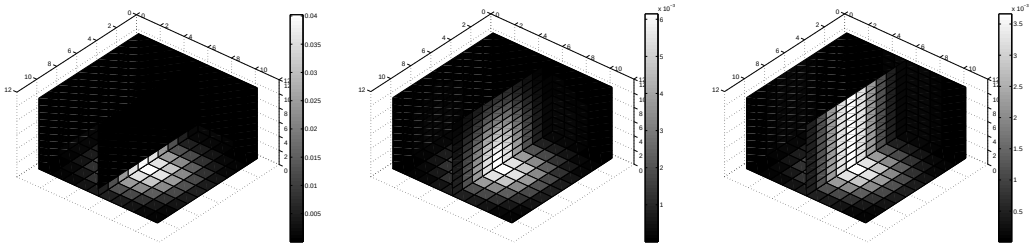
previous section, the nearest neighbor approach is to find the single block (C) that has the most similar precipitation to the previous day (B), and predict the precipitation amount (A) as being equal to that of the day after the nearest neighbor (D). The process is the same for multiple blocks in space and time, except that we search for the closest match to the entire block, using sum of squared error, treating a XxYxT block as a $X * Y * T$ dimensional vector. In other words, for every day (A) for which the precipitation level is queried, a XxYxT block (B) is identified, a corresponding match (C) is found using the nearest neighbor method and the precipitation on the day (D) succeeding the closest match (C) is assigned as the prediction. To measure the accuracy of the prediction, we find the sum of absolute error between this prediction and the real precipitation level given by the dataset.

4.2 Weighted nearest neighbor

Our next objective was to determine the variation in influence of nearby spatial and temporal elements when compared to those further away on the weather at any queried location. For this purpose, we use half of a three-dimensional Gaussian as weights, with the equation:

$$G(x, y, t) = exp(-\frac{x^2 + y^2}{2\sigma_{xy}^2} - \frac{(t - s_t/2)^2}{2\sigma_t^2}) \tag{1}$$

In our notation, s_x, s_y, and s_t are the latitude, longitude, and day dimensions of the block. There was not an obvious reason to differentiate between latitude and longitude, so we use the same standard deviation σ_{xy} for both. Similarly, σ_t defines how much previous days are taken into account. A block of size XxYxT is normalized to 1, which allows for straightforward comparison between blocks with different sigmas. As the σs approach infinity, this Gaussian weighted nearest neighbor approach becomes the same as the standard unweighted one. Figure 3 illustrates examples of the calculated weight.



(a) $\sigma_{xy} = 2$ and $\sigma_t = .1$ (b) $\sigma_{xy} = 2$ and $\sigma_t = 5$ (c) $\sigma_{xy} = 2$ and $\sigma_t = 100$

Figure 3: Planar slice views of three-dimensional Gaussian with different σ_ts

5 Experiments and results

We took test and training samples from the full 17x16x16801 (latitude X longitude X days) dataset. We tested block sizes from 1x1x1 to 9x9x9 for the unweighted algorithm, and used a 5x5x5 block with σs from .1 to 30 for the weighted algorithm. Unweighted results are shown in Figure 4 and weighted results in Figure 5.

5.1 Unweighted results

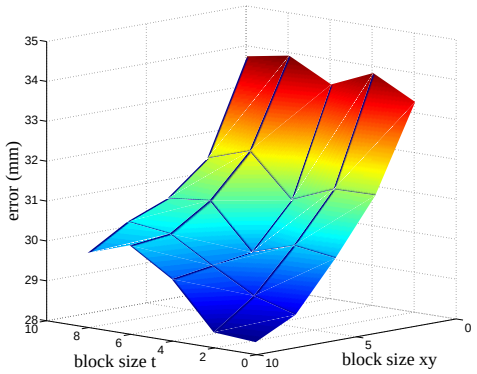


Figure 4: Results for unweighted nearest neighbor.

In the unweighted case, it is interesting to note that increasing the spatial search radius consistently increases accuracy, while increasing the temporal search radius generally makes things worse. This suggests that weather patterns shift much more dramatically and unpredictably over time than space. This is not terribly surprising, as intuitively, it seems more likely for the rain in a given location to be related to the rain a kilometer away, compared to the rain several days earlier in the same location.

5.2 Weighted results

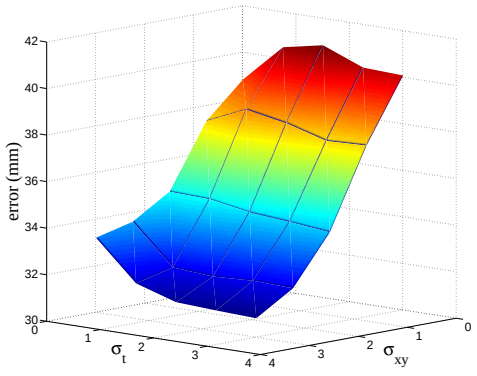


Figure 5: Results for weighted nearest neighbor.

For every sigma we tried, the weighted algorithm performed worse than the unweighted. In fact, the higher the sigmas used, and hence the closer the algorithm is to the unweighted case, the better it performs. This suggests that the amount of precipitation on a given day at a given time is not more correlated with that of nearby locations than with ones further away. This is understandable,

for several reasons. First, weather patterns can shift dramatically within even a single day, which is the temporal resolution of our data. In addition, global effects such as temperature are liable to be fairly constant across the entire spatial window, which could cause similar weather patterns across the entire grid. There may be correlations between particular locations, but they are not necessarily neighbors. Also, the source data does not include information regarding dramatic elevation changes over a small area [11]. For example, two mountainous regions might be separated by a valley and the precipitation levels in the mountainous regions will be similar to each other than that occurring in the valley, even though the valley is spatially closer to each of the mountains. We discuss possible changes to our algorithm that could explicitly model this effect in our future work section.

5.3 Implementation details

There are a few implementation details to consider. With data on a 17x16 grid, for 365 days a year for over forty years, doing a full nearest neighbor search of the entire dataset would take months of machine time. To speed up the process, we use a KD-tree that is created from the training examples, and search for the nearest neighbors for our testing examples. This makes it possible to use the entire dataset, although it was still too slow to be able to do a full N-fold cross validation. Instead, we took a random sample for training and test, and ran both the unweighted and weighted nearest neighbors with a variety of parameters. It might be possible to speed up the search if approximate-nearest neighbor is used, although it would still be necessary to compare the results with the complete nearest neighbor. In the case of the unweighted nearest neighbor experiment, the maximum block size used was 9x9x9. This constraint was necessary due to limitations in time and memory.

6 Future Work

There are a variety of possible extensions to our basic method. We can categorize them as either locally or globally based improvements.

6.1 Locally-based improvements

The most interesting improvement for the local features would be to try to discover which neighboring space-time locations affect a given location the most. This could be estimated using several different methods. The simplest would be to use a Naive Bayes approach, where each precipitation value at a specific x, y, t is used to independently predict the precipitation for a query location. The ones that are seen to be most effective would be given the largest weight. Alternately, logistic regression could be used in a similar manner. It is expected to capture relationships between blocks better than the independence assumption of Naive Bayes.

6.2 Globally-based improvements

At it's heart, weather is affected by events across the entire world. Any algorithm of high accuracy standards would need to take into account the weather in more than just the locations specified in the given dataset. However, there are interesting relationships that might be identified even with the current limited data. For instance, the data spans almost fifty years, and they are liable to inter-decadal climate variability. It would be interesting to try to fit the general trend to a model, and use the model to normalize the data to allow for more accurate matching.

6.3 Use of additional datasets

Deduction or availability of information on global phenomena, such as warm and cold fronts can improve accuracy of prediction by using this data as additional features. Semi-supervised learning techniques can be employed if local information is made available and is used appropriately. This information can be on terrain (whether it is a mountainous or coastal region), or partitioning the precipitation into rain vs. snow, and so on. Such data distinguishes the concerned locations from any other geographic location leading to wiser predictions.

An interesting aspect of weather prediction would be to assess which times of the year are easier to make weather predictions given a particular machine learning technique. Knowledge of this trend

will determine the accuracy of the predictions made. Such variabilities have been observed in [6], where simulated winter precipitation using general circulation models was generally less intense than that observed while summer predictions using the same models were quite similar to observed data.

7 Conclusion

The objective of this report is to predict precipitation levels based on a given dataset of precipitation levels across a 50 km x 50 km area and 45 years. The nearest neighbor method was used due to the complexity of the patterns in the data set, the limited availability of information in the form of features, and the computational limits of time and memory. We implemented both unweighted and weighted nearest neighbor. The unweighted method makes predictions based on closest matching space-time blocks and the weighted method performs such that it gave less importance to locations and days further away from the focus point in a Gaussian manner. With the unweighted algorithm, we found that accuracy in prediction increases with an increase in the spatial dimension and decreases with an increase in the temporal dimension. This conforms with the intuition that the weather in surrounding areas is more influential than the weather that occurred few days ago. The weighted nearest neighbor algorithm performed better with larger values of sigma. This implies that locations further away in space are as influential as those nearby. Based on our methods and findings, improvements in the algorithm, additional data, and other methods have been suggested. In conclusion, this report discusses the use of the nearest neighbor machine learning technique to predict the precipitation level with reasonable accuracy and more importantly, demonstrates the significance of different features used in prediction.

References

[1] Martin Beniston. *From Turbulence to Climate : Numerical Investigations of the Atmosphere with a Hierarchy of Models*. Springer, 1997.

[2] L. Bodri. Precipitation prediction with neural networks. *Acta Geodaetica et Geophysica Hungarica*, 36(2):207 – 216, 2001.

[3] NASA Ames Research Centre. Artificial intelligence research branch, 1991 progress report and future plans. Technical report, NASA Ames, 1991.

[4] M Ehrendorfer. Predicting the uncertainty of numerical weather forecasts: a review. In *Meteor. Z*, 1997.

[5] JNK Liu and RST Lee. Rainfall forecasting from multiple point sources using neuralnetworks. In *IEEE Systems, Man, and Cybernetics*, 1999.

[6] T.J. Osborn and M. Hulme. Evaluation of the european daily precipitation characteristics from the atmospheric model intercomparison project. *International Journal of Climatology*, 18:505–522, 1998.

[7] TN Palmer. Predicting uncertainty in forecasts of weather and climate. In *Rep. Prog. Phys.*, pages 71–116, 2000.

[8] Hans R. Pruppacher and James D. Klett. *Microphysics of Clouds and Precipitation*. Kluwer Academic Publishers, 1997.

[9] D Silverman and JA Dracup. Artificial neural networks and long-range precipitation prediction in california. *Journal of Applied Meteorology*, 39:57–66, 2000.

[10] DF Tucker and ER Reiter. Modeling heavy precipitation in complex terrain. *Meteorology and Atmospheric Physics*, 39(3-4):119–131, 1988.

[11] M. Widmann and C. S. Bretherton. Validation of mesoscale precipitation in the ncep reanalysis using a new gridcell dataset for the northwestern united states. *Journal of Climate*, 13:1936–1950, 1998.

[12] Wikipedia. Numerical weather prediction. November 28, 2006.