That's Hot: Predicting Daily Temperature for Different Locations

Alborz Bejnood, Max Chang, Edward Zhu
Stanford University
Computer Science 229: Machine Learning

December 14, 2012

1 Abstract

.
The problem of effectively modeling weather has been a focus of numeric simulations since the early 1950s. We address the specific task of modeling daily temperature using only temperature data from preceding days and equivalent days in previous years. Weather data compiled by the National Oceanic and Atmospheric Administration (NOAA) for five chosen cities over a period of 31 years was used as training data. An autoregressive analysis generated a temperature hypothesis function with mean absolute error of $2.58°F$ against the true temperature values. For validation of the applicability of a temperature-based data assumption, we used the k-nearest neighbors algorithm (with previous days' temperatures used as coordinates), obtaining a mean absolute error of $2.70°F$. The similar predictions of the two models suggests stronger confidence in the effectiveness of temperature modeling through a singular dependence on preceding temperature data.

2 Introduction

Weather simulations have predominantly studied weather changes on a relatively microscopic level, with less consideration of longer time scale data trends and regularities. This "small-change" assumption–that the weather tomorrow can best be determined by the weather from the preceding week–usually provides reasonably accurate short-term measurements (which is why it is often used for daily forecasts), yet performs poorly on predictions more than a week into the future due to the chaotic nature of atmospheric perturbations. Our approach targets a particular aspect of the weather-temperature-measured at noon daily at particular locations, and looks to generate a predictive temperature model using solely prior temperature information. To improve both the immediate and extended temperature predictions, we used weather data collected over a 30 year period as a training data set.

3 Data Acquisition

The data used was obtained from publicly available weather information from the Na-

tional Oceanic and Atmospheric Administration. Temperature data for various cities and locations, with information from over a hundred years ago to the current day, was obtained from airports (as weather-related information from airports are likely to be more reliable due to their relevance in flying) from the years 1982 to 2012. We decided (based on the data quality and availability) to analyze the temperature data in the following five locations:

1. London, United Kingdom
2. New York City, United States
3. Paris, France
4. San Francisco, United States
5. Tokyo, Japan

4 Temperature Models

We implemented two models: an autoregressive model to forecast the temperature one day ahead based on seven-day trailing temperatures, and a k-nearest neighbors model which looked at a weighted average of the 100 closest three-day trailing temperatures.

4.1 Autoregressive Model

An autoregressive model is a linear regression of the current value in a series against one or more prior values. More formally,

$$X_t = \theta_0 + \sum_{i=1}^{p} \theta_i X_{t-i} + A_t$$

where A_t is white noise, θ_i are the parameters, and X_t represents the value being predicted. In our case, the autoregressive model considered temperature values on a rolling 7-day basis, using those values to predict the temperature of the following day. The hypothesis function is given by

$$F_t = \theta_0 + \theta_1 F_{t-1} + \theta_2 F_{t-2} + \theta_3 F_{t-3} \quad (1)$$
$$+ \theta_4 M_{y-1} + \theta_5 M_{y-2} \quad (2)$$

where

t = time, in days
F_f = temperature as a function of t
y = time, in years

5 K Nearest Neighbors

The k nearest neighbors (KNN) algorithm is a method for supervised classification of an unknown object based on the closest examples of that object in the previously classified training data. The KNN process is as follows:

1. Determine the distance between the unknown object $\vec{x}_{in}$ and all examples $\vec{y}_{tr}$ in the training data, where the distance function $d(\vec{x}_{in}, \vec{y}_{tr})$ between the objects is given by

$$d(\vec{x}_{in}, \vec{y}_{tr}) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

2. Choose test examples corresponding with the $k(1 \leq k \leq n)$ smallest distances $d(\vec{x}_{in}, \vec{y}_{tr})$

3. Use a weighted average of the values in the training set to make a prediction for the unknown value

In the context of creating a temperature model, we let the temperatures of the days preceding the test day being predicted represent the coordinates of $\vec{y}_{tr}$ (such that $\vec{y}_{tr} = [T_{t-1}, T_{t-2}, ...T_{t-m}]$, where t is the day, T_t is the temperature in Fahrenheit, and m is the number of trailing days used in predicting T_t),

and found the k closest neighbors to $\vec{x}_{in}$. For example, if we look at the measured temperature in London over a 5-day span in August 2005, we might find that the temperatures are given by $\vec{y}_{tr} = [71°F, 78°F, 81°F, 74°F, 75°F]$. To predict the temperature on the following day, we look at all temperature vectors over 5 consecutive days, and calculate the distance (as defined above) between that vector and $\vec{y}_{tr}$. We then choose the 100 vectors corresponding with the minimum distances, check the following temperature value of each of the vectors in the training data set, and use a weighted mean of those known results to estimate $\vec{x}_{in}$. An empirical analysis suggested an optimal result of using the preceding three days, for which we let $k = 100$. Given the likelihood of periodic weather behavior on short-term scales (and correspondingly more arbitrary results when extrapolating farther backwards), we included an exponential decay weighting function to prioritize minimum-distance vectors occurring at a more recent time to the predicted day.

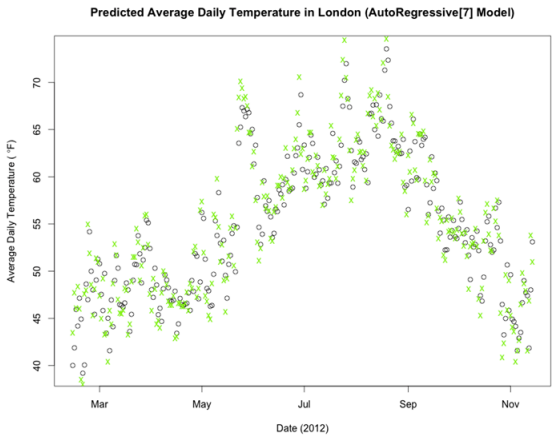

Figure 1: Autoregressive model for London in year 2012. The black o's represent the true data points, and the green x's represent the model's predicted values

6 Results

Figure 1 and 2 show the temperature function as predicted by the autoregressive and the weighted 100-nearest neighbors models for London, with the corresponding graphs for San Francisco shown in Figure 3 and 4. The mean absolute errors for the cities examined ranged from $1.66°F$ from the autoregressive model for San Francisco to $3.80°F$ from the KNN model in New York City (see Figure 5 for complete table of error values at all evaluated locations). In all cases, we note that the autoregressive model predictions slightly outperform those of the KNN model, though the difference in the errors remains relatively consistent. A possible explanation for this discrepancy would be due to the empirical determination of several of the parameters used. For instance, the size of the vector used in the KNN implementation (for which we chose to check the three prior days at each data point), as well as the number of closest such vectors considered (again chosen empirically with $k = 100$) are likely to have been suboptimal, despite relatively accurate results. Further testing would likely reveal stronger parameter values. Alternatively, the KNN algorithm might be improved in a multidimensional context, for which other factors besides only the temperature (such as wind speed, humidity, or periodic weather phenomenon such as El Nino) would also be included in an analysis. Similar arguments would apply to the autoregressive model; however, the coefficient values for more than three days before the actual day were found to be small (Figure 6), indicating that temperatures from more than three days prior to the current day have little predictive value. This point was confirmed using an Auto-Correlation Function (ACF). One par-

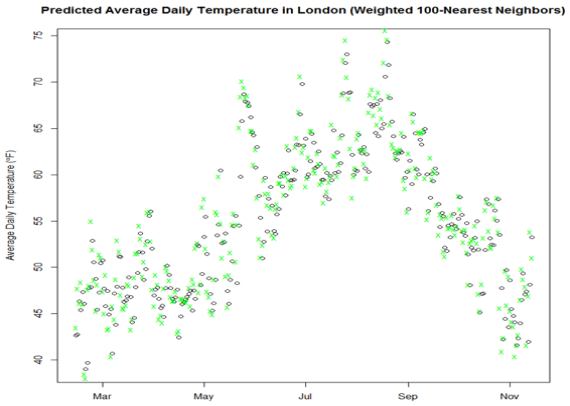

Figure 2: KNN model for London in year 2012. The black o's represent the true data points, and the green x's represent the model's predicted values

$$\tau(t) = \theta_1 \sin\left(\frac{2\pi}{365}t\right) + \theta_2 \cos\left(\frac{2\pi}{365}t\right) \quad (3)$$

$$+\theta_3 \sin\left(\frac{2\pi}{7}t\right) + \theta_4 \cos\left(\frac{2\pi}{7}t\right) \quad (4)$$

$$+\theta_5 t + \theta_6 \quad (5)$$

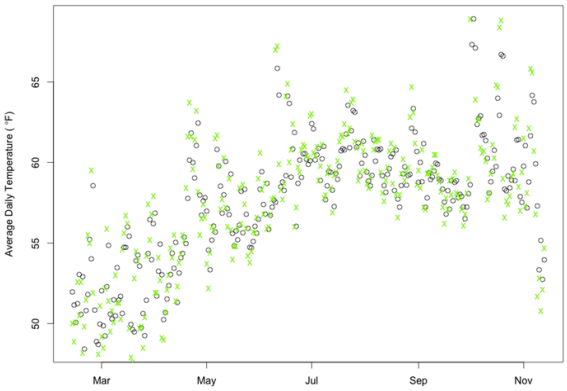

Figure 3: Autoregressive model for San Francisco in year 2012. The black o's represent the true data points, and the green x's represent the model's predicted values

ticular observation with regards to the error shows a link between a region's climate and the predictive value of the models used. Intuitively, one expects that a geography with less temperature variance would correspond with more accurate predictions, as increased temperature ranges introduce greater fluctuations that are difficult to accurately anticipate. The results in Figure 5 are consistent with that logic, given that San Francisco relatively consistent temperature patterns has a mean absolute error of more than $2°F$ less than that of New York City, which undergoes a much larger range of temperatures over the course of a year. Several efforts to generate more realistic models produced less accurate results. One such model hypothesized that the temperature function could be assumed to be periodic over the course of a year, with a hypothesis function defined by

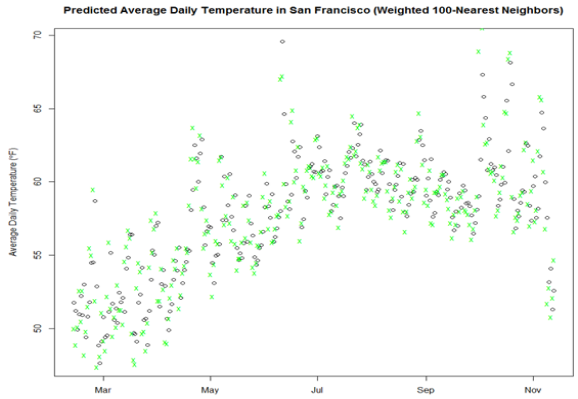

Figure 4: KNN model for San Francisco in year 2012. The black o's represent the true data points, and the green x's represent the model's predicted values

City	Algorithm	Mean Absolute Error	Standard Error of Absolute Error	Sum of Squared Errors
NYC	k-NN	3.802721	3.076586	6570.203
NYC	AR(7)	3.549901	2.818757	5642.532
Tokyo	k-NN	2.533271	2.119811	2996.048
Tokyo	AR(7)	2.441286	1.999954	2734.916
SF	k-NN	1.750536	1.681403	1617.333
SF	AR(7)	1.667284	1.55894	1430.255
London	k-NN	2.486743	2.025215	2824.38
London	AR(7)	2.435809	1.886866	2607.133
Paris	k-NN	2.937818	2.22237	3726.731
Paris	AR(7)	2.803379	2.096085	3365.046

Figure 5: Table of cities analyzed, with corresponding algorithms used and error values

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.478806   0.259935  13.383  < 2e-16 ***
previous1    0.067275   0.009514   7.071 1.63e-12 ***
previous2    0.027611   0.012741   2.167   0.0303 *
previous3    0.027898   0.012861   2.169   0.0301 *
previous4    0.028773   0.012885   2.233   0.0256 *
previous5    0.084472   0.012863   6.567 5.36e-11 ***
previous6   -0.185903   0.012744 -14.587  < 2e-16 ***
previous7    0.888745   0.009520  93.359  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Coefficients associated with 7-day autoregressive analysis

7 Conclusions

The results support the idea that a sophisticated model to predict temperature can provide accurate results through use of only a temperature data set. The autoregressive model and the KNN model both produced predictions to within a mean absolute error of $4°F$ of the true temperatures. Future analysis would be liable to include the following:

- Use the above method on all locations (instead of a select few cases) to increase potential to notice trends, in turn improving parameter estimations to reduce overall error
- Use other, more sophisticated techniques to verify/refine algorithmic results
- Include knowledge about other factors (such as El Nino effect, hurricanes, etc) to provide a more comprehensive, multi-dimensional analysis
- Extend this method to other locations with poorer data quality to check for consistency

8 References

1. IURL: ftp://ftp.ncdc.noaa.gov/pub/data/gsod/
2. Importing Weather Data from Wunderground — California Soil Research Lab, http://casoilresource.lawr.ucdavis.edu/drupal/node/99 10 November, 2012
3. Hayati, M et all. "Temperature Forecasting Based on Neural Network Approach." World Applied Sciences Journal, 2007.
4. Baboo, S et all. "An Efficient Weather Forecasting System using Artificial Neural Network." International Journal of Environmental Science and Development, Vol 1. October 2010.
5. Lai, L et all. "Intelligent Weather Forecast." International Conference on Machine Learning and Cybernetics, Shanghai. August 2004.