Weather forcasting in Alaska using machine learning

Osuke Sashida
Computer Science(Data Science)
University of Alaska Fairbanks
Fairnanks, United States
osashida@alaska.edu

Abstract

Weather forecasting is an important part for people to live safely or effectively. Especially to people who work in the area deeply related with nature such as agriculture, ocean fisherman, ski resort managers, etc. Because weather has big impact for such area. For example, fisherman can't go ocean in the rain. So, people have gathered weather data for predicting future weather. And it can be feasible using machine learning. In this paper, I use weather data and create linear regression model and polynomial regression model to predict future weather situation.

Keywords—subject area, technology used, etc.

Weather forecasting, Machine learning, linear regression, polynomial regression

I. INTRODUCTION (SHOULD NOT GO BEYOND THIS 1ST PAGE)

As I mentioned in abstract, to solve this problem leads to many rewards in jobs related with nature. For instance, people who work in such fields can avoid bad effects caused by weather. And I have two reasons that I chose this topic. First, I thought I can make sense about machine learning through weather forecasting. Because I often check the weather in the morning to live comfortable throughout a day. That's why it is intuitive data for me. Second, I believe most people in Alaska need this machine learning model. The weather in Alaska is hard for people especially in winter. Besides, I'd like to make people more comfortable by my product. Therefore, this topic is suitable for me.

In this paper, I use the data[1] including "Date", "Wind", "Solar", "Elevation", etc. at some place from 1979 to 2014. This dataset format is CSV. We can precisely predict weather using this dataset. Because this dataset accumulates a lot of information related with weather and geographic.

I already have this data set as CSV. Hence, I change it to pandas data frame and eliminate abnormal data such as row lacking some part to analyze easily. Next, I separate by location data and calculate average value that relate with weather situation such as max temperature for each month. In this paper, I focus on temperature data mainly. And I calculate average value for each month. Then I visualize data to confirm weather change for each month and decide how to make a model from this data. Basically, I create linear regression model and polynomial regression model and compare these models. Finally, I predict temperature using my models and visualize to make it useful and helpful.

In many case, linear regression models return the lower average error(separated by month) than polynomial regressions. Besides, they execute totally different prediction for the data. And error values widely spreads and depends on when month is. In addition, in most month, 2 degree polynomial regression model perform the best.

II. BACKGROUND (

III. MUST FINISH IN 1-COLUMN)

A. NumPy

This is the fundamental or core library of all other python library. If developer uses Numpy instead of default python list, they can save memory and run code faster. In addition, we can write code more easily. Numpy has a lot of convenient functions. For example, it has broadcasting, some mathematical operation and function related with date data. So, I'd like to write efficiently and quickly.

B. Pandas

Pandas is an open-source python package that makes users easy to handle data sets including two-dimensional data. We can execute complex data operation such as grouping and finding null and statistical calculation easily using pandas. If we execute these operations in excel, it is hard. That's why I use pandas to find some pattern from the data set. In addition, it is interchangeable with the other library.

C. MatplotLib

We can analyze data sets statistically or effectively or comprehensively using matplotlib. Because it makes using it easy to visualize data as various types of figures like histograms, plot charts and so on. In other words, we can get visualization data whatever we want. In this paper, I try to analyze time series separated by location. So, it is important for such analytics to visualize information. It means I can get validation for my analytics or model by visualizing.

D. Scikit-learn

This is open-source machine learning library that supports supervised and unsupervised learning. It also gives developer some useful tools that necessary in the process of machine learning such as model fitting, data processing, model evaluation and so on. I'd like to predict future weather from this dataset. Hence, I need to use machine learning through scikit-learn.

IV. Data Description (0.75-1 column)

The data that I use in this paper are given by the tutor in this course. This data set is accumulated as CSV file. It has 10 columns including date data, location data, and weather situation data and more than four hundred thousand rows. Fortunately, this data set doesn't include null value. This data set has 35 different location weather data if I separate by "Longitude" and "Latitude". Almost data observed every day from 1979 to 2014. The standard deviation of maximum temperature, minimum temperature and solar is higher than other columns. The standard deviation of precipitation and wind and relative humidity is low value. I'll mainly focus on the column which has high standard deviation. Because it means its data change widely.

V. Methodology

Brief overview

At first, I read the CSV raw data and convert to pandas data frame. For this data, I don't need to do data cleaning. Because this data doesn't have any error or null values. Then I apply scaling to weather data except location data. And I separate data into by month and location. By doing this, I can calculate average temperature of each year for each location. In addition, I separate these data into train data and valid data. I make linear regression model and polynomial regression model for the combination of each location and each month. I predict max temperature in 2015 to 2030 that the raw data doesn't include.

A. Linear regression model

I used the linear regression model in scikit learn. And I used the data from 1979 to 2004 as training data and the data from 2005 to 2014 as valid data. I made linear regression model fit the training data and predict the max temperature of validation data and also future duration. I select the mean squared error (MSE) to calculate error of each model.

B. Polynomial regression model

The trick to make good polynomial regression model is to choose degree correctly. That's why this model create from 2 to 10-degree model for each combination of location and month. Then it chooses the one degree model that has the lowest MSE value. And I use the linear regression model in scikit learn. So, its method is similar to linear regression.

VI. Results and Discussion

I calculated the average error, average data variance in each month and what degree the polynomial regression model chose. Then I show this result as TABLE 1. As you can see the TABLE1, we can say linear regression models are superior to polynomial regression models in terms of error. In addition, polynomial regression models selected 2-degree polynomial in most case. And I convert the TABLE.1 data to Fig.1 graph. Fig.1 describes variance values highly relate with linear regression model's errors. But it shows variance values a little relate with polynomial regression model errors.

$$E(y_0 - \hat{f}(x_o))^2 = Var\left(\hat{f}(x_o)\right) + [Bias(\hat{f}(x_o))]^2 + Var(\varepsilon)$$

And basically, following formula underlie regression model. Expected value of the difference between real data and expected value can be expressed like that. In this formula, I suppose $y_0 = f(x_o) + \varepsilon$. And I also suppose error is distributed like $\varepsilon \sim N(0,1)$. Essentially, high degree models have high flexibility, and it leads to low bias. However, it leads to high variance simultaneously. Conversely, linear model (degree = 1) have low flexibility and it leads to high bias. But it causes low variance. We call these relationships bias-variance tradeoff. Therefore, we need to select the model appropriately. And I think the reason why the models choose two-degree polynomial regression is caused by these relationships. But I can't find why linear regression models superior to polynomial regression models and even two-degree models. The weather fact that only weather specialist can find underlie this data. In the future, I'd like to work on this fact with weather specialist to predict weather situation completely.

Month	*Sum error (Linear)*	*Sum error (Polynomial)*	*Average Variance*	*Degree (Polynomial)*
1	1.506	1.535	0.0438	{2:35}
2	1.515	1.678	0.0420	{2: 24,3: 6,10: 5 }
3	1.970	1.804	0.5017	{2: 33, 10: 2}
4	2.036	1.810	0.0400	{2: 34, 10: 1}
5	0.835	0.735	0.0196	{2: 32, 3: 3}
6	1.587	1.588	0.0314	{2: 34, 10: 1}
7	1.589	1.426	0.0247	{2: 24, 3: 4,9: 1, 10: 6}
8	0.950	1.003	0.0255	{2: 35}
9	0.987	1.046	0.0243	{2: 35}
10	0.958	1.444	0.0295	{2: 5, 3: 21, 9: 1, 10: 8}
11	1.958	4.489	0.0544	{ 2: 2, 3: 33}
12	1.350	1.537	0.0393	{2: 35}

TABLE I. The values in each month

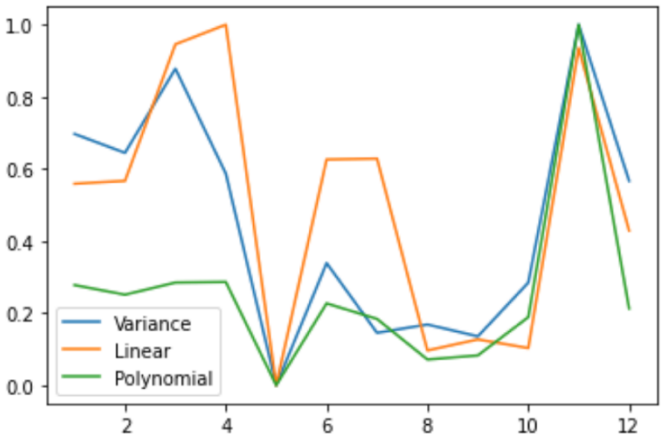

Fig.1. Scaled linear regression error and polynomial regression error and avarage variance for each month

ACKNOWLEDGMENT

Professor Arghya Das

REFERENCES

[1] https://docs.google.com/spreadsheets/d/e/2PACX-1vQfwFLOECjiMWuBd6Zu1GCpQIPJiRd73w-0rVAYZE4XfvBwmKdcN6JfLRdJP63kMUUeeHRD806uQe0z/pub?output=csv

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.