Understanding the Weather Data to model rainfall Prediction in a region

Debatri Mitra
Electrical Engineering and Computer Science
University Of California Irvine
debatrim@uci.edu

1 Introduction

This projects looks into the dataset of rainfall prediction based on various temporal features coming from different sensors which consist of information of different weather condition like air velocity, direction, cloud density, humidity etc. The task is to predict the amount of rainfall in a region given the sensor data.
In order to achieve this goal a systematic sequential statistical approach has been used. Firstly, the data is analyzed to understand the importance of different features along with identifying the redundant or less contributing features

2 Learning from the Data

This article uses and investigates a Kaggle Data-set courtesy of UC Irvine's Center for Hydrometeorology and Remote Sensing, including Dr. Soroosh Sorooshian, Dr. Xiaogang Gao, Dr. Kuo-lin Hsu, Dan Braithwaite, and Yumeng Tau, and additionally processed by Jonathan Stroud and Nick Gallo (ICS). The data contains many temporal features obtained using the geo-sensors around the point of interest along with image patches around the point of interest. However, in this part we explore only the temporal features to learn the model. There are 91 temporal features which includes standard statistical features like various moments of cloud velocity, density,surface pressure etc.

2.1 Scoring Metric

We pose it as a regression problem where the task is to predict the amount it would rain under the given location and available data. The scoring is based on Root Mean Squared Error (RMSE) given in eq (1)

$$MSE = (\frac{1}{m} \sum (\hat{y}^i - y^i)^2)^{\frac{1}{2}} \tag{1}$$

The use l_2 norm based error metric is pretty standard in regression problem and often is a good idea to use. Participants with lower score are ranked higher naturally.

3 Understand the Features

In order to understand the importance of different features it is very important to understand the quality of the features or in other words the relative importance of the features should be investigated before proceeding. The first step in any machine learning task is to see the correlation between different features in order to be able to discard highly correlated features as in most cases they don't contribute much in the inference and repetitive features should be discarded as a first step of dimensionality reduction. Fig. 1 shows the correlation matrix between the features.

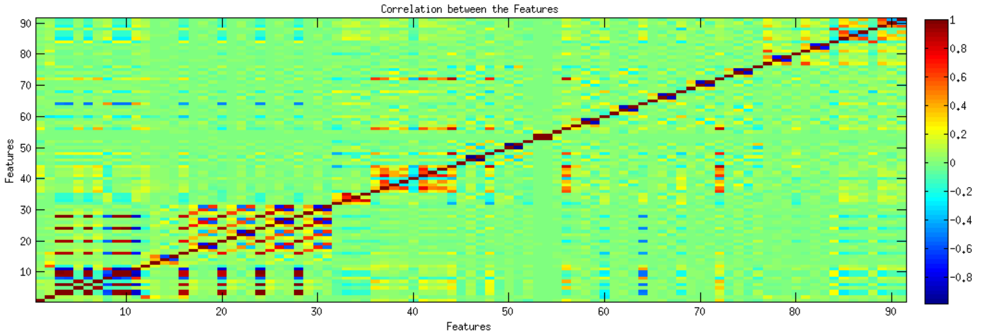

Figure 1: Correlation between temporal Features

Looking at the correlation matrix it can be seen easily that there are not too many highly correlated features as the number of features with correlation¿0.8 was less than 1.5%. Thus it can be safely assumed that the data was very nice and features are mostly wisely chosen. If two features are highly correlated only one has been kept.

3.1 A very simple Model to Start with

To start with and get a feel of the data a linear regression model has been fit a linear regression model has been trained with all the 91 available features as it is. The score obtained is 0.62135

3.2 Bayesian MARS model

The result obtained from simple linear regression urges to investigates if the nonlinear model is a better fit to the dataset. Here a Bayesian Multivariate Adaptive Regression Spline (MARS) model has been trained on the dataset which is basically an extension of the linear model that automatically models non-linearities and interactions between variables. It builds a model of the form (2)

$$\hat{f}(x) = \sum_{i=1}^{k} (c_i B_i(x)) \tag{2}$$

where c is constant co-efficients and B(x) are the basis functions. Here each basis function can be either a constant, a hinge or a product of multiple hinge functions. MARS automatically selects variables and values of those variables for knots of the hinge functions.
Training with MARS model leads to a slight improvement and reduces the error to 0.62135.

3.3 Decision Tree

The next model that has been tried is the decision tree. There are multiple parameters in the decision tree like minimum number of parents, depth etc which can be tuned and have effect in the prediction. However there is a trade off in the model selection as increasing the depth and minParent parameters improve the prediction on training it has overfitting effects thus choosing proper values of these parameters is very important.
Starting off with a simple decision tree, in order to understand the effect of depth cut-off the max Depth has been varied between 1-15 and the model performances are compared using a 5-fold cross validation with a split of 70-30 training and test data. The performances are compared in (2) From the comparison it seems like for this data-set maximum depth of 8-10 works best giving least validation error. For depth 10 the score on the competition turns out to be 0.65843.

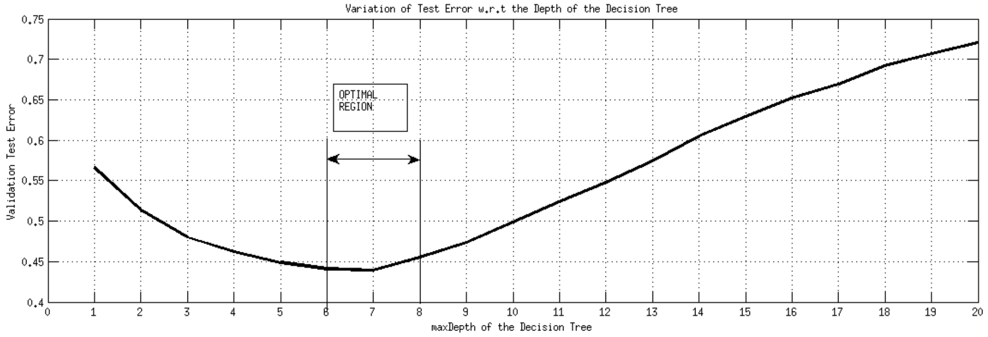

Figure 2: Comparing performance at different depth

In order to understand the role of min number of parent it was necessary to do a similar performance comparison between different values of minimum parent parameter. Variation of Parent parameter keeping depth constant shows that it performs better with the cut-off of 800-1000 for this data-set as shown in (3). When combined the two optimized parameter we obtain a score of 0.62135 which is quite improved than the previous one.

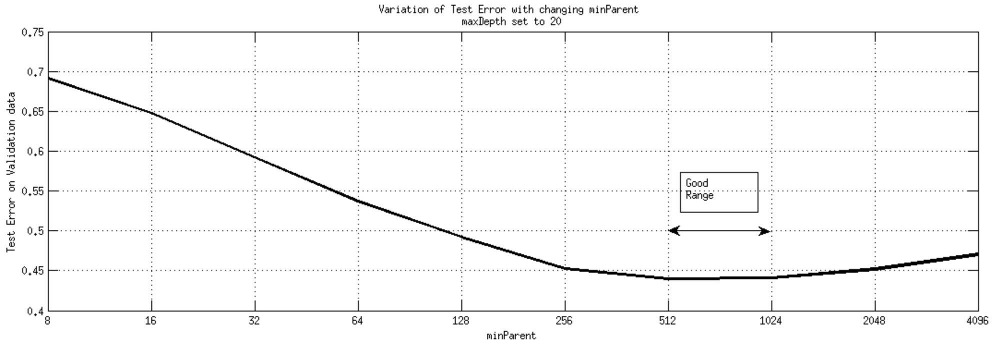

Figure 3: Comparing performance at different number of parent

3.4 Random Forest

Ensemble based methods are strong and often leads to good results. Random forest is one of the most popular ensemble methods used is Random Forest which is basically bagged decision tree. Here a Random Forest model has been constructed by assembling 25 decision trees where each decision tree is trained over 60 randomly chosen features and restricting to a maximum depth of 10. The score obtained is 0.609 and better than the other two models trained on. The results are optimized in terms of depth selection as we selected the depth giving best result of the data-set to be the depth of each of the learners.

4 Dimensionality Reduction

As we notice from the consecutive learners tried on this dataset the improvement has been only by a small fraction it can be inferred that dimensionality reduction based approach might be useful. Also, we see that some of the features are correlated with each other, and thus reducing their number might be helpful.

Thus I performed PCA on the data. The explanation of variance from the principal components is shown in the Figure 4. We see that around 95% of the variance is explained by the first 50 principal components. The last 21 components have no contribution in the variance of the data. I chose to use the first 70 principal components to retrain the random forest. Setting the other principal components as zero, I transformed the current features to the principal component space, and retrained the random forest.
Optimal choice of number of Principal Components is a key here. So, we split the training data into a 80-20 validation set where the 20% is used as test data, and the rest as the training dataset. We trained a random forest for the data transformed with the top 10, 20, 30, 40, 50, 60, 70, 80, and 90 principal components. Results are described in Table 1. Table 2 shows the Variance explained - Clearly, 40 components performed the best with the given random forest. Thus, the data has around 40 degrees of freedom. Therefore, we use the

Number of PC	Score
10	1.08
20	0.96
30	0.81
40	0.63
50	0.68
60	0.68
70	0.71
80	0.78
90	0.83

Number of PC	Variance Explained
1	12.08
10	48.13
20	66.63
30	79.9
40	89.67
50	96.4
60	99.59
70	100

first 40 components to train the final model. I chose an ensemble of 25 trees for this forest. After training the model, I transformed the test data to this new space, and obtained the prediction values. The final test score of this model was 0.608. This is the my best performing model so far, and shows an improvement of 0.1 over the previous best result. The top performing model has an MSE of 0.581, which is fairly close to my best performance.

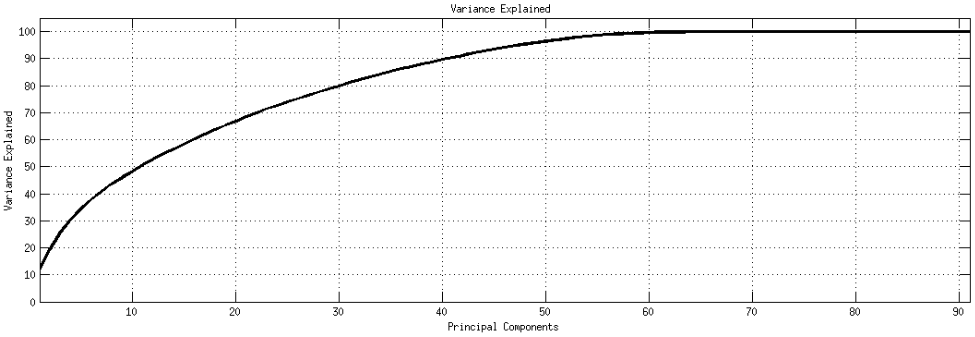

Figure 4: Variance Explained by different Principal Components

5 Conclusion

The problem I tried to solve was, to predict the amount of rainfall based on some weather parameters like wind speed, direction, temperature etc. The metric used was Mean Squared Error.

To study the data, I first studied the correlation between the features of my data. I found that most features did not correlate with each other. However, 1.5% of the entries in the correlation matrix had correlation higher than 0.8. I kept only one of those features which correlated highly with each other.

To start, I used a multivariate linear regression model, which performed fine. To improve my results, I tried Multivariate Adaptive Regression Spline model, which improved the results significantly. To further improve my results, I used a decision tree. I experimented with the tree's depth and the pruning condition using minimum number of points at each parent. This improved the results slightly.

Finally, I tried to use an ensemble. I trained a random forest, with 25 trees like the ones described above. The results slightly improved. Finally, to improve my results even more, I used dimensionality reduction on the data by performing PCA. After experimenting with the number of principal components to use, I decided on using the first 40 principle components. This gave me my best results, suggesting that the data had some reduncdant features, which only contributed to noise. Myfinal result was not very far from the best result in the contest. However, I think, my results can be improved by further experimentation with the random forest training parameters.

6 Reference

- Quinlan, J. R. (1987). ”Simplifying decision trees”. International Journal of Man-Machine Studies 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.

- Y. Yuan and M.J. Shaw, Induction of fuzzy decision trees. Fuzzy Sets and Systems 69 (1995), pp. 125139
- Deng,H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).
- Shaw P.J.A. (2003) Multivariate statistics for the Environmental Sciences, Hodder-Arnold.
- Barnett, T. P., and R. Preisendorfer. (1987). ”Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis.”. Monthly Weather Review 115.
- Hsu, Daniel, Sham M. Kakade, and Tong Zhang (2008). ”A spectral algorithm for learning hidden markov models.”. arXiv preprint arXiv:0811.4413.