HOW MUCH DID IT RAIN ?

By,

SAI SUCHITH MAHAJAN

Gmail : mahajan.saisuchith@gmail.com

Overview

Today water is an essential commodity, for growing food, providing drinking water, and countless other uses. As climate change progresses, past historical records of rainfall totals may not be altogether great estimates of future rainfall total. For agriculture also it is extremely important to know how much it rained on particular field. One option is direct measurement of rainfall total via rain gauges.However rainfall varies with space and time so it's impossible to have rain gauges everywhere.Therefore, remote sensing instruments such as radar are used to provide wide spatial coverage. The machine learning problem, then, is given radar measurements for a location and time, such as radar reflectivity, precipitation type, and average precipitation shape, predict a probability density of how much rain fell. This is the challenge that we address here.

Unlike a conventional Doppler radar, a polarimetric radar transmits radio wave pulses that have both horizontal and vertical orientations. Because rain drops become flatter as they increase in size and because ice crystals tend to be elongated vertically, whereas liquid droplets tend to be

flattened, it is possible to infer the size of rain drops and the type of hydrometeor from the differential reflectivity of the two orientations.

We are given polarimetric radar values and derived quantities at as location over the period of one hour. You will need to produce a probabilistic distribution of the hourly rain gauge total

DATA

Data is available on : https://www.kaggle.com/c/how-much-did-it-rain/data

The data set consists of 1,126,694 training points. Each data point is a collection of numerical radar features collected in one hour for some particular location. These locations vary amongst several midwestern states in the United States, and were taken between April and November 2013. The data set provides no location or time features, and were shuffled, so that there is no immediate way to recover location or time features. There are 19 provided features, with three of these features being rain rates predicted from three current algorithms. These three past algorithm features, RR1, RR2, and RR3, are respectively, the 'HCA-based', 'Zdr-based', and 'Kdp-based' algorithms. The other 16 features are given as time series numerical data. An example data point could have its 'TimeToEnd' feature s '58.0 55.0 52.0 49.0 41.0,' indicating radar information taken at 58, 55, . . . , 41 minutes from the end of the hour. For this same row, the features 'Reflectivity' as '0.0, 0.0, 1.2, 4.5, 0.0' and 'RR1' as '0.0. 0.0, 2.2, 0.3, 0.0' mean these measurements taken at the time points in the 'TimeToEnd' series. The label for each row is one float number, the amount in mm of rain collected for that hour. The test set consists of 630,521 points. The test set draws from the same collection of radars covering the same region, but in the next year, 2014. It is not indicated whether the test points were drawn according to the same time 2 or location distribution as in the training set for 2013. For the online Kaggle competition, submissions are predictions of the probabilistic distribution of the hourly rain total. Each row of the submission is a list of values $P(y \leq Y)$, for Y integer values 0, 1, 2, . . . 69, and y the rainfall total, in mm.

SIMPLE BENCHMARK MODELS

Inspection of a histogram of the training set labels gives us that about 87.64% of the time, the rain label was 0.0 mm. This makes intuitive sense, in that most of the time, for most regions of the United States, it is not raining, under the assumption that the training samples were not favorably drawn from rainy periods or locations. Hence we tried a 'No Rain' null-hypothesis prediction on the test data consisting of all 1 predictions, i.e. the predicted probability that it rained less than any amount was always 1. Somewhat surprisingly this classifier was competitive. With a score of 0.01017651, this submission ranks tied between places 207 through 220 on the online leaderboard.

The another simple benchmark model was calculating the proportions of labels in the train set and predicting same proportion in the test set. This was also competitive with a score of 0.00971225 and submission rank will be near to 180.

MODELING

I transformed the model into classification model which has 70 classes and each class represent will represent the value lying between (i,i+1) and 'i' starts from 0 to 68. I also made another class which contains only '0' as expected rainfall. So total will be 70 classes. And lastly I removed all the train examples which have expected rainfall greater than 69.

FEATURE SELECTION

In the model I have included 6 features " RR1, RR2, Radar quality Index, Number of scans ,Reflectivity quality control and hybrid scan." I have replaced all the missing data with 0.

I have used XGB classifier to train the model and also tried to change the parameters like learning rate but the score got worsened. So I used all the default parameters.

With this model the submission rank will be 31 in private leaderboard and score of 0.00782772 in private leaderboard.

ACKNOWLEDGMENTS

I would like to thank everyone who helped me and special thanks to kaggle forums and kaggle blogs and my friends.

REFERENCES

https://www.kaggle.com/c/how-much-did-it-rain/discussion

http://blog.kaggle.com/2015/07/01/how-much-did-it-rain-winners-interview-1st-place-devin-anzelmo/

http://blog.kaggle.com/2015/06/08/how-much-did-it-rain-winners-interview-2nd-place-no-rain-no-gain/