ECE 5984 SP22 – Prof. Jones – Group Project I

Due Thursday, March 22, 2022 – 11:59 PM via Canvas

In this project, your team will develop a pair of models to predict the next day's precipitation at a specific weather station, using current weather data. You will write Python code to load and prepare the data, to reformat the data to be suitable for modeling, to train both a decision tree model and a linear regression model, and to compute performance for each model. Part of your assignment is to experiment with parameters and settings to achieve the best performance possible.

Data:

You will obtain the data from the US National Oceanic and Atmospheric Administration, at the following URL (daily weather by station). https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_station/ . Each file represents a set of historical daily data from a single weather station. Each Project team will use a different data set, as shown in the following table:

Team	File to use
Team A	USW00024018.csv
Team B	USC00129430.csv
Team C	USW00013891.csv
Team D	USW00094849.csv
Team E	USW00023066.csv
Team F	USW00024128.csv
Team G	USW00014840.csv
Team H	USW00094014.csv
Team I	USW00014848.csv
Team J	USW00093820.csv
Team K	USW00013880.csv
Team L	USW00026617.csv
Team M	USW00014936.csv
Team N	USW00014898.csv

The data contains a list of date-stamped readings, such as max temperature (TMAX). You can find a version of the data documentation at the following URL, though there are some inaccuracies, so be careful: https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt

The data is a compressed comma-separated value (csv) file; you will need to uncompress the data for use.

Data Preparation:

As a csv file, the data does not have any column names in the first row. You will need to read in the data and assign appropriate names.

You should perform the usual steps of assessing the data quality, fixing any aberrant values and so on. Note carefully the units used in the data fields, as noted in the documentation!

To perform modeling, this data must be transformed into a dataset containing one row per day, with columns for all available readings on each day. In other words, on the row for Jan 1 2000, I would like to have columns for precipitation on that day, snow on that day, etc. Once this is done, run another quality report to check for any problems, and address them if they exist.

You must also create a pair of target columns. First, create columns for PRECIPFLAG and PRECIPAMT. PRECIPFLAG is set if there was any precipitation (rain, snow, etc) on the day in question. PRECIPAMT is the total amount that occurred, in inches of rainfall (assume that eight inches of snow is equivalent to one inch of rain). Now, create a pair of columns for NEXTDAYPRECIPFLAG and NEXTDAYPRECIPAMT, which are the values of the flag and amount variables for the next day in the set. These will be your two target variables.

Compute and add to your report the frequency of occurrence and/or the histogram (as appropriate) of the target variables.

Normalize the data appropriately. Divide the dataset into training and test partitions, using a 70/30 split. Use a random seed so that your code runs repeatably.

Modeling - Classification:

From this data set, develop and test two sets of predictive models. First, a decision tree classifier to predict the occurrence of precipitation on the next day (the NEXTDAYPRECIPFLAG target). Use both entropy and Gini criteria and see which gives the best performance on test (not used in training) data. Measure the performance as the True Positive Rate:

$$TPR = \frac{|TestSetCorrectlyPredicted|}{|TestSet|}$$

Experiment with different tree depths, and any other parameters that you wish, to obtain the best performance. Record everything you tried and the corresponding performance achieved. For your final model, save the tree shape as a .png file. Also, make a list of the three or four most important variables (first used in the tree).

Modeling – Regression:

Develop and test a linear regression predictive model to predict the next day's precipitation amount. The target value will be the continuous variable NEXTDAYPRECIPAMT. Measure the performance as the Mean Square Error:

$$MSE = \frac{1}{|TestSet|} \sum_{1}^{|TestSet|} (NEXTDAYPRECIPAMT - MODELOUTPUT)^2$$

You should try two forms of regression models: a LinearRegression and a RidgeCV model. Experiment with different options and parameter settings if appropriate. Record everything you

tried and the corresponding performances achieved. Also, make a list of the three or four most important variables (having the largest magnitude of coefficient).

Use of Prior Days' Data:

So far, your models have been trained using one day's data. In other words, you are attempting to predict the weather on Jan 3 from data collected on Jan 2. Now, I want you to modify the data set to include two days' weather data as features; we want to predict weather on Jan 3 using data from both Jan 1 and Jan 2.

To do this, modify the final preprocessed data frame to duplicate the columns from each day as new, additional columns on the next day's row. We are going to take the predictor columns from Jan 1, and add them (under new names) to the data from Jan 2.

Once you have done this, retrain your best performing decision tree model and your best performing linear regression model. Measure the impact on performance, and present it clearly in a single table. Also, be sure to save the decision tree as a .png file.

Report:

Your report shall contain:

- A title
- Team names
- An abstract
- Information on the station: filename, location, range of dates in the data, etc.
- Data quality reports on the raw data from the file and on the final preprocessed data just prior to partition for modeling
- All results on your decision tree models, as described above (including pictures of each of the decision tree models)
- All results on your linear regression models, as described above
- All results on the models after including prior days' data
- Discussion
- Conclusions
- Appendix – your Python code pasted in as plain text

Use the IEEE Transactions report template, attached to this assignment.