Assignment 2: Data Preparation

In this assignment, you will practice how to prepare datasets for analysis. As real-world data usually comes with a variety of formats and might contain errors, we will use data cleaning method to prepare in to analysis. In this assignment, you will get your hands on a set of hourly weather data, and would be asked to load the dataset, apply data cleaning methods, and train a precipitation forecast model.

This homework is due on **March 4th, at 11:59 PM EST**.

Please upload the submission as a single .zip file to CMS. It should include:

1. A report that includes the description of the full process of your data preparation and model training, and the evaluation of your forecast model.
2. The source code for all your experiments and the instructions on how to run your code to reproduce the results claimed in your report.

You can perform the following tasks using Python and any additional libraries that you wish to use. PANDAS is our recommended package for handling data.

Introduction

The file `Hourly_Weather_Surface_Brazil_2_Cities.zip`[1] contains hourly weather data from 2 weathers stations of a southeast region in Brazil. It includes 17 climate parameters from 2 weather stations, which should be comprehensive enough to perform precipitation forecast (rain prediction). However, since the weather stations are not always in working order and the data recording is not rigorous enough, there are a lot of missing fields, missing entries and obvious errors.

Your job in this assignment is to select **one city** of your choice, clean up the data and train a precipitation forecast model which takes history climate parameters as input and predicts the amount of precipitation in the upcoming hours.

Data Cleaning

One problem of the given data is the mixed use of zeros (0) and blank entry. For solar radiation (gbrd) column, while almost half of the entries are missing, common sense suggests that all the missing entries happen after sunset, when there is no sun at all (thus zero solar radiation). Thus in this scenario, imputation by constant values (0) is appropriate. Use data imputation, list-wise deletion and/or other methods of your choice to further clean the data. You only need to clean the features that you need for the next step.

1 This is based on a subset of “Hourly Weather Surface - Brazil (Southeast region)” dataset: https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region

Outlier Removal by Z-Score

Due to some technical issues, there is a small portion of data containing erroneous records in air pressure related fields (stp, smax, smin). Use the z-score method to identify and remove those obvious outliers.

Precipitation Forecast

You will build a model that predict the amount of precipitation for the next hour, given the data of past 3 hours. You may choose whatever set of features you want and prepare the dataset accordingly.

Linear regression model alone may not work well for this task as there is a huge gap between rain and no rain. A better idea is to train two models:

- A binary logistic regression model that predicts whether it will rain in the next hour.
- A linear regression model predicting how much it will rain.

Report the detailed data cleaning procedures you use. Please provide reason for your every major design decision, such as why you perform certain data cleanings, and why you use certain parameters. Train the same model on both minimally cleaned dataset that has missing fields filled with zero and thoroughly cleaned dataset and report the performance of both models using appropriate metrics. How do removing outliers affects performance? Explain Why.

Bonus

As an optional opportunity for those who want to receive a small extra credit, finish one or two of the following tasks:

- Extend your model to do longer term (5 hour, 10 hour, 1-5 hours, etc) precipitation forecast.
- Evaluate your prediction model on the data of another city. Does the model trained for one city transfer well to the other city?

Appendix

Explanations of data fields in the provided data:

wsid	Weather station id
wsnm	Name station (usually city location or nickname)
elvt	Elevation
lat	Latitude
lon	Longitude
inme	Station number (INMET number) for the location
city	City
prov	State (Province)
mdct	Observation Datetime (complete date: date + time)
date	Date of observation
yr	The year (2000-2016)
mo	The month (0-12)
da	The day (0-31)
hr	The hour (0-23)
prcp	Amount of precipitation in millimetres (last hour)
stp	Air pressure for the hour in hPa to tenths (instant)
smax	Maximum air pressure for the last hour in hPa to tenths
smin	Minimum air pressure for the last hour in hPa to tenths
gbrd	Solar radiation KJ/m2
temp	Air temperature (instant) in celsius degrees
dewp	Dew point temperature (instant) in celsius degrees
tmax	Maximum temperature for the last hour in celsius degrees
dmax	Maximum dew point temperature for the last hour in celsius degrees
tmin	Minimum temperature for the last hour in celsius degrees
dmin	Minimum dew point temperature for the last hour in celsius degrees
hmdy	Relative humid in % (instant)
hmax	Maximum relative humid temperature for the last hour in %
hmin	Minimum relative humid temperature for the last hour in %
wdsp	Wind speed in metres per second
wdct	Wind direction in radius degrees (0-360)
gust	Wind gust in metres per second