MSAN 621 Introduction to Machine Learning Fall 2016 Module 2

Assignment 1

There are two portions to this assignment: a practical portion and an analysis portion. For the practical portion, you will use two machine learning algorithms against two data sets. Once complete, you will answer the questions below based on the results you see from applying the algorithms against the data in the practical portion. You will submit your implementation code for the practical portion and your answers (as a text file or document in Word, OpenOffice or PDF) for the analysis portion.

Practical Portion

For the practical portion, you will use the two machine learning algorithms, *Linear Regression* and *k Nearest Neighbours*, against two data sets. You may [read: should] use these algorithms as implemented in scikit-learn. The implementations should have exactly one output: MSE (mean squared error) on the test set. Your implementation should take the test set (eg. name(s) of the test weather stations — see below) as a parameter.

There are two data sets for this assignment: the Boston data set and the U.S. Monthly Climate Normals data set.

The Boston Data Set

The Boston data set is available at

`https://github.com/dbrizan/MSAN621-data/blob/master/boston.csv`

This file is in CSV format with the first line as the header.

You will train your models with all data except the last 50 rows, which are reserved to test your model's performance. Your model will infer the final column (Median value of owner-occupied homes in $1000's) based on the features from the other columns.

You will report MSE (mean squared error) in the analysis portion below. Ensure that your implementation outputs only this MSE number.

The U.S. Monthly Climate Normals Data Set

The U.S. Monthly Climate Normals data set is available through data.gov, specifically at the URL:

`ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/`

In the subdirectory path `products > hourly`, the file `hly-temp-normal.txt` contains the 20-year mean temperature for each of the 457 weather stations in the USA. Each line of the file is space-separated, formatted as shown in Table 1.

Note the following about the temperature data found in fields 4-27:

1) Most temperature readings are followed by a character (eg. "C" for "Confirmed"). You should ignore this trailing character.
2) A number of temperature readings are unavailable or missing. These are indicated with the value "-9999" or similar flag values.

Field #	***Description***
1	ID of the weather station
2	Month of collection (01-12)
3	Day of the month of collection (01-31, varying by month)
4-27	Hourly temperature at station for hours: 0:00 (midnight) - 11:00PM

Table 1: Fields in hly-temp-normal.txt file

For the missing values, you should use the scikit-learn Imputer class to infer values of the missing values. You may use either the mean or median strategy.

You will build your models with most of the data (not including the material from the weather stations used to test your model). You will test the models with some of the data, specifically from the following five (5) weather stations: USW00023234, USW00014918, USW00012919, USW00013743 and USW00025309. The features for your models will be as follows:

- The previous hour's temperature reading from the same station, if available
- The previous day's temperature reading at the same hour from the same station, if available
- The mean temperature of that hour's reading (across all stations) on that day
- The mean temperature for that day up to, but not including, the hour in question

You will report MSE (mean squared error) for each of the five test weather stations in aggregate and individually — i.e. six total results per model — in the analysis portion below, but ensure that your implementation outputs only this MSE number.

Analysis Portion

Based on the results of your implementation above, answer the following questions:

1. Which Imputer class strategy did you use for missing values?
2. What, if anything, can be done to improve the missing temperature readings?
3. What are the performances of your models against the data in terms of MSE?
4. In what ways does this performance align with your expectations or fail to do so?
5. In what ways would you explore improving the performance of the models?

Grading

The implementation portion of this assignment is worth 80% of the grade, and the analysis portion the balance, evenly divided among the questions. The implementation is pass/fail (i.e. 100% or 0%) based on the specification above.

Assignments may be submitted up to 14 days late for 80% credit.