CSE 351 - Introduction to Data Science (Spring 2022)

Assignment 2: Prediction/Modelling

Deadline: April 10, 2022 11:59 PM EST.

Data

The goal of this homework is to develop a method to predict the electricity usage based on the weather conditions. We provide the following two datasets for this task:

1. **Weather**: Weather data for one year with daily weather conditions
2. **Energy Usage**: Energy usage history for one year (in kW) with 30-minute intervals. The energy usage of specific devices like AC, Fridge, washer, etc. are also given.

You will need to submit your code (programs/source files) in three different formats (.ipynb, .pdf and .py). Make sure that you properly document your program (code) with proper comments highlighting the exact sequence of operations which are required to arrive at the resulting tables and figures. The submission instructions are provided at the end of the assignment.

Tasks

1. Examine the data, parse the time fields wherever necessary. Take the sum of the energy usage (Use [kW]) to get per day usage and merge it with weather data **(10 Points).**

2. Split the data obtained from step 1, into training and testing sets. The aim is to predict the usage for each day in the month of December using the weather data, so split accordingly. The usage as per devices should be dropped, only the "use [kW]" column is to be used for prediction from the dataset **(5 points).**

3. Linear Regression - Predicting Energy Usage:

Set up a simple linear regression model to train, and then predict energy usage for each day in the month of December using features from weather data (Note that you need to drop the "use [kW]" column in the test set first). How well/badly does the model work? (Evaluate the correctness of your predictions based on the original "use [kW]" column). Calculate the Root mean squared error of your model.

Finally generate a csv dump of the predicted values. Format of csv: Two columns, first should be the date and second should be the predicted value. **(20 points)**

4. Logistic Regression - Temperature classification:

Using only weather data we want to classify if the temperature is high or low. Let's assume temperature greater than or equal to 35 is 'high' and below 35 is 'low'. Set up a logistic regression model to classify the temperature for each day in the month of December. Calculate the F1 score for the model.

Finally generate a csv dump of the classification (1 for high, 0 for low)

Format: Two columns, first should be the date and second should be the classification (1/0).

(20 points)

5. Energy usage data Analysis:

We want to analyze how different devices are being used in different times of the day.

- Is the washer being used only during the day?

- During what time of the day is AC used most?

There are a number of questions that can be asked.

For simplicity, let's divide a day in two parts:

- Day: 6AM - 7PM

- Night: 7PM - 6AM

Analyze the usage of any two devices of your choice during the 'day' and 'night'. Plot these trends. Explain your findings. **(10 points)**

6. Visual Appeal and Layout - For all the tasks above, please include an explanation wherever asked and make sure that your procedure is documented (suitable comments) as good as you can.

Don't forget to label all plots and include legends wherever necessary as this is key to making good visualizations! Ensure that the plots are visible enough by playing with size parameters. Be sure to use appropriate color schemes wherever possible to maximize the ease of understandability. Everything must be laid out in a python notebook (.ipynb). **(5 Point)**

Submission

1. This assignment must be done individually by every student. Your code will be checked thoroughly to detect copying/plagiarism. Do your own work!

2. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!

3. Please use Piazza to ask any questions.

4. Submit everything through Blackboard. You will need to upload:

a. The Jupyter notebook all your work is in (.ipynb file)

b. Python file (export the notebook as .py)

c. PDF (export the notebook as a pdf file)

d. Linear regression and logistic regression csv dumps

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse351_hw2_lastname_firstname_sbuid.ipynb
2. cse351_hw2_lastname_firstname_sbuid.py
3. cse351_hw2_lastname_firstname_sbuid.pdf
4. cse351_hw2_lastname_firstname_sbuid_linear_regression.csv
5. cse351_hw2_lastname_firstname_sbuid_logistic_regression.csv