STA 4364 HW 2

due September 26 by 11:59PM

Submission Format: Please submit your homework as 1) a HTML or pdf document, and 2) also submit the source file in either R Markdown or Jupyter notebook format (at most one of each type of file).

Problems can be done in Python or R. ISL = Introduction to Statistical Learning textbook.

Problem 1: (7 points) ISL Chapter 3 Problem 9. This problem involves linear regression of the Auto dataset from the last homework. The dataset is available on Webcourses in the file `Auto.csv`. This dataset is referenced in the text of both Chapters 2 and 3 and you can look at the information there as a guideline. One important thing to note is that you should omit rows with missing entries using the `na.omit()` function in R or the or the `df.dropna()` function in `pandas`, where `df` is your dataframe. (*Note: In general, omitting NA values might bias your results. But data imputation can be involved so for now we will just remove missing entries*).

Problem 2: (7 points) Analyze the Air Quality dataset provided in Webcourses (file is `AirQuality.csv`). This data was collected hourly in Beijing from January 1, 2010 to December 31, 2014. We are interested in PM2.5 concentration as the response variable. The dataset attributes are:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- pm2.5: PM2.5 concentration (ug/m^3)
- DEWP: Dew Point
- TEMP: Temperature
- PRES: Pressure
- cbwd: Combined wind direction (cv is SW)
- Iws: Cumulated wind speed (m/s)
- Is: Cumulated hours of snow
- Ir: Cumulated hours of rain

Analyze the data using the following steps:

(a) Prepare the dataset: load the data, omit rows with NA entries. Plot the histogram of PM2.5. Since this variable is clustered around 0, it's best to take a log transformation to make the data more symmetric and unimodal. Remove 0 entries for PM2.5 to avoid $-\infty$ values when doing the log transform (there are only 2 observations with PM2.5 of 0 so this is not a problem. In general care should be taken with 0 entries when doing log transforms to avoid taking the log of 0).

(b) Do an exploratory analysis on the data. Make plots or perform calculations to investigate the following questions:

- Does pollution appear to be increasing, decreasing, or neither over time?
- Are certain months, days of the month, or hours of the day, associated with greater pollution?
- Are environmental factors associated with greater or less pollution?
- Are there relations between the environmental factors?

(c) Perform a transformation on month, day, and hour that is appropriate for linear regression with cyclic variables:

$$x \mapsto (\cos(2\pi x/\tau), \sin(2\pi x/\tau))$$

where $\tau = 12$, 30, or 24 respectively. Remove the original coding of the cyclic variables from your dataset. Then split the data into a training and validation set.

(d) Create a linear model predicting PM2.5 using all other features (including No, how should this coefficient be interpreted?). What do you notice about the significance of the coefficients? Why is this happening? Are these significance values meaningful?

(e) Narrow down the predictors to a group of about 4 to 6 predictors that are most meaningful. Possible tools for doing this are doing univariate regression with each predictor, or multivariate regression with smaller number of observations selected randomly, or forward/backward regression with increasing or decreasing numbers of coefficients (see ISL Section 3.2.2), or a combination of these methods. Do a multivariate regression on PM2.5 using your smaller set of predictors. How does the R^2 value compare to the full regression from the previous part?

(f) Calculate the mean-square-error (MSE) on your validation set for a model with all predictors and a model with your reduced set of chosen predictors:

$$\mathrm{MSE}(\hat{\beta}) = \frac{1}{n_{\mathrm{val}}} \sum_{(X_i, Y_i) \in \mathcal{D}_{\mathrm{val}}} (Y_i - X_i^{\intercal} \hat{\beta})^2$$

where $\mathcal{D}_{\mathrm{val}}$ represents the pairs (X_i, Y_i) in your validation set and n_{val} is the number of validation observations. Interpret your model by revisiting the questions in Part (b).