STT 811

Homework 6

Due April 5, at 11:59:59 pm

1. ISLR 8.11 (using xgboost). Note that this problem can be done in Python if preferred.
    a. Part (a)
    b. Part (b)
    c. Part (c)
    d. Perform a grid search for the optimal hyperparameters in the model. Let nrounds go from 50 to 550 in steps of 100, maxdepth from 1 to 2, and eta be 0.01, .1, and .2.

2. ISLR2 9.7
    a. Part (a)
    b. Part (b) (with a 80/20 train/test split)
    c. Part (c)

3. Take a look at the nndb dataset available in the sample data. We will focus on the columns for protein, fat, sugar, carb, and fiber for this assignment.

(Remember to standardize the data)

    a. Perform k-means clustering for these fields with 4 and 8 clusters using the algorithm from scratch (not using the kmeans command). Repeat a few times; do you get the same cluster centers?
    b. Next, use k-means clustering with the kmeans command. Try from 2 to 10 clusters. For how many of these cluster numbers do you get the same centers after repeating kmeans (for each number do 3 iterations of kmeans)?
    c. Take the highest number of clusters for which kmeans gives a consistent response (same centers after repeating). Look at the food names for the data in each cluster. Can you give a verbal description of each cluster, based on the names?
    d. Perform a linear regression, predicting calories based on these 5 inputs for the entire dataset. Then do separate regression models for each of the clusters in (c), keeping only significant terms for each. How do the models compare (accuracy, adjusted R^2, etc.)?

4. For the gas data in the oil gas dataset
    a. Perform trend seasonality decomposition, comparing additive and multiplicative. Which one look better.
    b. Create forecasts for the gas data, with
        i. naïve
        ii. seasonal naïve
        iii. simple exponential smooth
        iv. Holt
        v. Holt-winters
    c. Calculate the MAPE's for each of the models in (b). Which fits the best?
    d. Create a plot of the best forecast along with the fitted values to show the fit.