ECE 5984 SP22 – Prof. Jones – HW 4

Due Thursday, April 5, 2022 – 11:59 PM via Canvas

PART 1 (15 points)

For this part, you are to explore the effect of k on the nearest neighbor classification algorithm. In the Datasets section of the Files page on Canvas, there is a file called "BattingSalaries.xlsx". This will be the source dataset. You are to implement a Python program to do the following:

1. Load the dataset, and perform missing value processing as usual.
2. Consider only the data from the year 2016.
3. Use each of the numeric features as predictors (no text or categorical fields).
4. Divide the datasets into training and test sets, using the value 22222 as the random seed value.
5. For values of k from 1 to 15:
    a. Implement a k nearest neighbor classifier to predict the binary variable 'lgID'
    b. Train the classifier on the training set
    c. Measure the accuracy on the test set, using the score method of the classifier.
6. For values of k from 1 to 15:
    a. Implement a k nearest neighbor classifier to predict the multiclass variable 'teamID'
    b. Train the classifier on the training set
    c. Measure the accuracy on the test set, using the score method of the classifier.
7. Create a plot of accuracy versus k for both classifiers. I want to see both plots on the same chart; use primary and secondary axes since the levels of accuracy are different.
8. Choose the best value of k for each problem (binary and multiclass classification). Discuss your choice. Are they the same? Why or why not?
9. Repeat the process using a different random seed value. Are the best choices for k the same as before? Why or why not?

PART 2 (10 points)

For this part, you are to compare the mean square errors for multivariate linear regression models on the Batting Salaries dataset and on yearly partitions of it. Once again, "BattingSalaries.xlsx" will be the source dataset. You are to implement a Python program to do the following:

1. Load the dataset, and perform missing value processing as usual.
2. Use each of the numeric features as predictors – except for Salary, of course.
3. Also use one-hot encoding to derive features from the league ID and team ID fields. (no text or categorical fields).
4. Implement a linear regression model to predict the continuous variable 'Salary'.
    a. Divide the dataset into training and test sets, using the value 22222 as the random seed value;
    b. Train the regressor on the entire training set;
    c. Measure the accuracy on the test set, as both the r^2 value and the mean square error.
5. For each year present in the dataset:

a. Write out the year and the size of the dataset for just that year;
b. Divide the dataset for just that year into training and test sets, using the value 22222 as the random seed value;
c. Implement a linear regression model to predict the continuous variable 'Salary';
d. Train the regressor on that year's training set;
e. Measure the accuracy on that year's test set, as both the r^2 value and the mean square error.

6. Plot the MSE versus year, to determine which year's salary is most "predictable"

SUBMISSION

Assemble a single Word or pdf file containing:

- Your code for both parts 1 and 2; please paste as plain text with no dark mode;
- Your graphs for items 7 and 9 (two curves on each) of part 1
- Your discussion for items 8 and 9 of part 1
- Your graph for item 6 of part 2

Please submit your Word or pdf file via Canvas. Also attach your .py file(s); if you use Jupyter, please export and attach a .py file (don't submit your .ipnyb notebook file). Don't put your files into a zip archive; submit individual files.

Notes:

- I wrote the output to the console, pasted into Excel and used Excel to create my charts. You can do it this way or some other way.
- All processing of the file must be done by your Python code. You are NOT to edit the BattingSalaries.xlsx file in any way!
- This may be helpful to you: https://pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html