

Portfolio Assignment: ML with sklearn

Objectives:

- Gain experience with machine learning in sklearn on a small data set
- Compare experience in R versus sklearn

Turn in:

- Use Google Colab or a local Jupyter notebook, go to File Print and print to pdf
- Upload the pdf to eLearning
- Upload the pdf to your Portfolio

Instructions:

1. Read the Auto data (5 points)
 - a. use pandas to read the data
 - b. output the first few rows
 - c. output the dimensions of the data
2. Data exploration with code (5 points)
 - a. use describe() on the mpg, weight, and year columns
 - b. write comments indicating the range and average of each column
3. Explore data types (5 points)
 - a. check the data types of all columns
 - b. change the cylinders column to categorical (use cat.codes)
 - c. change the origin column to categorical (don't use cat.codes)
 - d. verify the changes with the dtypes attribute
4. Deal with NAs (5 points)
 - a. delete rows with NAs
 - b. output the new dimensions
5. Modify columns (10 points)
 - a. make a new column, mpg_high, and make it categorical:
 - i. the column == 1 if mpg > average mpg, else == 0
 - b. delete the mpg and name columns (delete mpg so the algorithm doesn't just learn to predict mpg_high from mpg)
 - c. output the first few rows of the modified data frame
6. Data exploration with graphs (15 points)
 - a. seaborn catplot on the mpg_high column
 - b. seaborn relplot with horsepower on the x axis, weight on the y axis, setting hue or style to mpg_high
 - c. seaborn boxplot with mpg_high on the x axis and weight on the y axis
 - d. for each graph, write a comment indicating one thing you learned about the data from the graph

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.

7. Train/test split (5 points)
 - a. 80/20
 - b. use seed 1234 so we all get the same results
 - c. train /test X data frames consists of all remaining columns except mpg_high
 - d. output the dimensions of train and test
8. Logistic Regression (10 points)
 - a. train a logistic regression model using solver lbfgs
 - b. test and evaluate
 - c. print metrics using the classification report
9. Decision Tree (10 points)
 - a. train a decision tree
 - b. test and evaluate
 - c. print the classification report metrics
 - d. plot the tree (optional, see: <https://scikit-learn.org/stable/modules/tree.html>)
10. Neural Network (15 points)
 - a. train a neural network, choosing a network topology of your choice
 - b. test and evaluate
 - c. train a second network with a different topology and different settings
 - d. test and evaluate
 - e. compare the two models and why you think the performance was same/different
11. Analysis (15 points)
 - a. which algorithm performed better?
 - b. compare accuracy, recall and precision metrics by class
 - c. give your analysis of why the better-performing algorithm might have outperformed the other
 - d. write a couple of sentences comparing your experiences using R versus sklearn. Feel free to express strong preferences.