COMP40370 Practical 3

REGRESSION AND CLASSIFICATION

Prof. Tahar KECHADI

Assignment Files

- `./Practical-03-A.pdf   Assignment questions (this file).`
- `./auto-mpg.csv         Data file.`

Expected output files

- `./Prcatical-03.ipynb       Python notebook solutions.`
- `./Prcatical-03.html        Python notebook in HTML format.`

Requirements: Python 3.9+, pandas 1.3+, sklearn 0.24+.

Part A: Regression Analysis

This practical work aims to use regression to predict a continuous-valued variable. The dataset for this practical describes the fuel consumption of some cars, the same dataset used in Practical 01. The dataset file is called "`auto-mpg.data`".

Question 1: Polynomial Regression

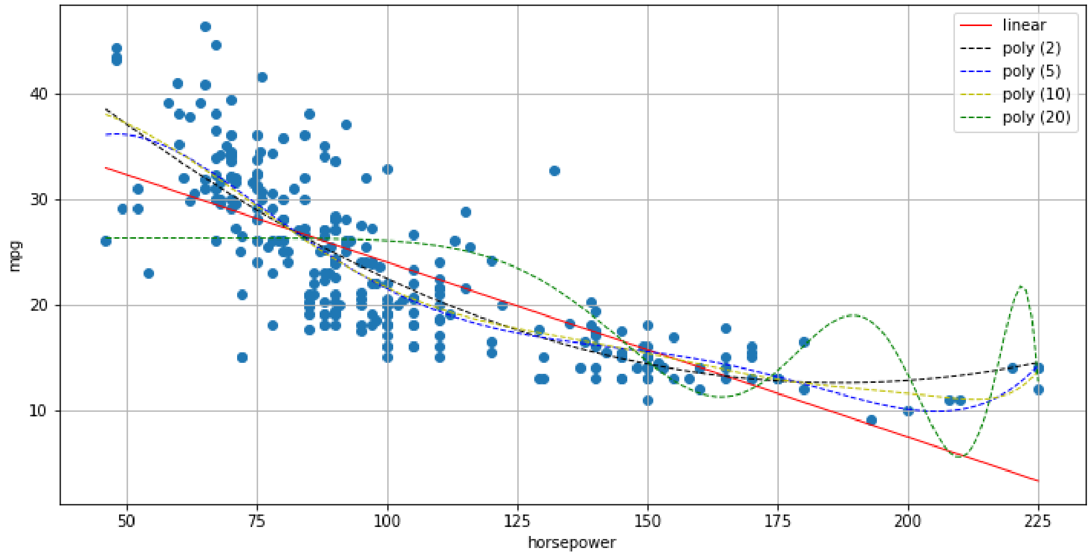
a. according to the data pre-processing instructions in Practical 01:
    1. Fill in the missing values, replace the outliers and remove duplicates rows of the dataset.
    2. Replace the number of column 'origin' by the country of origin as follows: 1:'USA',2:'Europe',3:'Japan'
    3. Create a new column 'brand', and extract the car brand from 'car name' column, replace any misspelling as follows: 'chevroelt': 'chevrolet', 'chevy': 'chevrolet', 'vokswagen': 'volkswagen', 'vw': 'volkswagen', 'hi': 'harvester', 'maxda': 'mazda', 'toyouta': 'toyota', 'mercedes-benz': 'mercedes'. Feel free to pre-process it as you see fit.

b. Generate a simple linear regression model that predicts `mpg` based on 'horsepower' alone (use `sklearn.linear_model` library) and train it with 70% of the data. What are the RMSE and R^2 values of testing/predicted data?

c. Generate a group of polynomial models (use `sklearn.preprocessing.PolynomialFeatures`) that predict `mpg` based on the 'horsepower' alone. The models should have degrees of 2, 5, 10, and 20. For each model, find RMSE and R^2 values of testing/predicted data.

d. Plot a scatter diagram between `mpg` and 'horsepower' with all the fitted linear and non-linear models. Explain what will happen when the model complexity is increased. Which model is better? Explain your answer.

An **illustration** of scatter diagram:



Question 2: Multiple Linear Regression

a. Predict mpg based on 'horsepower', 'displacement', 'weight' and 'acceleration'. What are the RMSE and R^2 values? Normalize the data then make the prediction again, What are the RMSE and R^2 values after normalization?
b. Which are the most influential two factors in the mpg prediction?
c. Predict mpg based on the first PCA component of the above four variables ('horsepower', 'displacement', 'weight' and 'acceleration'). Do you see any accuracy reduction?
d. Add origin and cylinders as categorical variables appropriately into the prediction model and discuss any accuracy changes.
e. By adding "model_year" as an ordinal variable and "brand" as a categorical variable do you see any improvement in the model performance?

Part B: Classification – Decision Trees

1. Create a new column from pre-processed data-frame in Q1 (a) called 'FEGroup' to categorise cars as "*high-fuel*" and "*low-fuel*" consumption. Allocate 10% of cars having the lowest mpg (use pandas *qcut* function) into the '*high-fuel*' consumption class. What is the base accuracy? The same question for the case where all cars are classified as 'low-fuel'.

2. Using *sklearn.tree* library, generate a decision tree based on 'horsepower', 'weight', 'acceleration', 'cylinders' and 'origin'. Using information gain (entropy) as the splitting criterion, set *max_depth* to 3 while leaving everything else to their default values.

3. Find the training and test accuracies when the data is split into 70% for training and 30% for testing. (***Hint:*** use *np.random.seed(42)*). Plot the resulting decision tree using sklearn *plot_tree* function.

4. Train other decision trees, but this time with *max_depth* set to 3, 5, 8, respectively. Compare the results of these 3 DTs with the previous model. Discuss the problem of measuring accuracy in an imbalance class problem, such as this one.

Please make sure that you have completed this practical. Next week, you will get the second part of the practical.