Lab Course Machine Learning
Exercise 2

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 08.11.2019 LearnWeb 3115

November 2, 2019

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. You should submit a zip or a tar file containing two things a) python scripts and b) a pdf document.
2. In the pdf document you will explain your approach (i.e. how you solved a given problem), and present your results in form of graphs and tables.
3. The submission should be made before the deadline, only through learnweb.
4. Unless explicitly mentioned, you are not allowed to use scikit, sklearn or any other library for solve any part. All implementations must be done yourself.

1 Exercise Sheet 2

1.1 Pandas (10 Points)

- **Dataset Exploration**: Download the dataset files that are provided on learnweb. The 1985 Auto Imports Database is made up of information about the cars that were imported during the year 1987.
  - **import-85.names** is the helper file which contains information regarding the dataset
  - **import-85.data** is a CSV file with the actual data

  You will analyze this dataset using pandas library and plot some interesting information using matplotlib library.

- Load the data using pandas
- Find the mean, median and standard deviation for each NUMERIC Column
- Doing the above is obviously not the most ideal method since it gives no real information.
- Group data by the field 'make'.
  * Find the average **price** , average **highway-mpg** and average **city-mpg** for each make.
  * Use a seaborn pairplot to visualize all int64 data types. Explain the plot, what information can we take out of it.
  * Similar to the first exercise use city-mpg as your dependant variable and engine-size as the independent value. Fit a line, use scatterplot for the data points and plot the line you predicted on top.
  * Comment on the fit and explain if it is a good prediction? if not, why?

1.2 Linear Regression via Normal Equations (10 Points)

In this exercise you will implement (multiple) linear regression using Normal Equations. See lecture (slides: 2-15) (Click here to download lecture).The learning algorithm is given on the slide 8.

- Reuse dataset from Excercise 1. Load it as X_{data}, [Hint:] from loaded data you need to separate ydata i.e. Price, which is your target.
- Choose those columns, which can help you in prediction i.e. contain some useful information. You can drop irrelevant columns. Give reason for choosing or dropping any column.
- Split your dataset X_{data}, Y_{data} into $X_{train}, Y_{train} and X_{test}, Y_{test}$ i.e. you can randomly assign 80% of the data to a X_{train}, Y_{train} set and remaining 20% to a X_{test}, y_{test} set.
- Implement learn-linreg-NormEq algorithm and learn a parameter vector β using Xtrain set. You have to learn a model to predict sales price of cars i.e. , ytest.
- Line 6, in learn-linreg-NormEq uses SOLVE-SLE. You have to replace SOLVE-SLE with following options. For each option you will learn a separate set of parameters. (Implement this yourself)
  - (a) Gaussian elimination
  - (c) QR decomposition
- Perform prediction $\bar{y}$ on test dataset i.e. X_{test} using the set of parameters learned in steps 5 and 6 (Hint. you will have two different prediction models based on the replacement function from step 6).

- Final step is to find how close these two models are to the original values.
  - plot residual $\epsilon = |y_{test} - \bar{y}|$ vs true value of y_{test} for each model.
  - Find the average residual $\epsilon = |y_{test} - \bar{y}|$ of each model.
  - Find the root-mean-square error $(RMSE) = \sqrt{\frac{\sum_{n=1}^{N}(y_{test}(n) - \bar{y}(n))^2)}{N}}$ of each model.

1.3 ANNEX

- You can use numpy or scipy in build methods for doing linear algebra operations
- You can use pandas to read and processing data
- You can use matplotlib for plotting.
- You should not use any machine learning library for solving the problem i.e. scikit-learn etc. If you use them you will not get any points for the task.