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: 16.11.2018 LearnWeb 3112

November 10, 2018

1 Exercise Sheet 2

1.1 Pandas (10 Points)

- **Dataset Exploration**: Download Gasprices.csv Click here to download Data File[1]. This dataset contains information about the sales of gas stations across a city along with other attributes. You will analyze this dataset using pandas library and plot some interesting information using matplotlib library.
    - Load the data using pandas
    - Summarize each NUMERIC field in the data, i.e. mean, average etc.
    - Group data by the field 'Name'.
        - Find the average price,average income and average number of pumps for each group.
        - Use a boxplot that visualizes the statistical information about (price, pumps, gasoline).
        - Use the Price and Income features in order to plot a prediction line similar to the first exercise. Normalize the Income (implement this yourself) and plot the line again. Comment on the different of the two plots.

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 9.

- Reuse dataset from Excercise 1. Load it as X_{data}, [Hint:] from loaded data you need to separate ydata i.e. Income, 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 houses 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
  - (b) Cholesky decomposition
  - (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 three different prediction models based on the replacement function from step 6).

- Final step is to find how close these three 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 $(RMSE) = \sqrt{\frac{\sum_{n=1}^{N}(y_{test}(n) - \bar{y}(n))^2)}{N}}$ 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.

[1] Data taken from James Scott