Lab Course Machine Learning
Exercise 3

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 015.11.2019 LearnWeb 3115

November 9, 2019

Instructions

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

1. You should submit two things a) python scripts(zipped) / jupyter notebook 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 3

1.1 Data preprocessing (5 Points)

1.1.1 Datasets

Airfare and demand: target $->$ price
Wine Quality: target $->$ quality
Parkisons Dataset: target $->$ total_UPDRS

You are required to pre-process the datasets.

1. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use hot-one encoding. [Hint: use hashmap (dict) or pandas.get_dummies]. Please explain your solution.

2. If required drop out the rows with missing values or NA. In next lectures we will handle sparse data, which will allow us to use records with missing values.

3. Split the dataset into 80% Train set and 20% Test set.

4. Are there any columns that can be dropped? if so, which ones are why.

1.2 Linear Regression with Gradient Descent (15 Points)

You are required to present your results on all three datasets.

Part A: (8 Points): Implement Linear Regression with Gradient Descent

In this part you are required to implement linear regression algorithm with gradient descent algorithm. Reference lecture https://www.ismll.uni-hildesheim.de/lehre/ml-19w/script/ml-02-A1-linear-regression.pdf

For each dataset given above

- 1. A set of training data $D_{train} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\}$, where $x \in R^M, y \in R$, N is number of training examples and M is number of features
- Linear Regression model is given as $\hat{y}^n = \sum_{m=1}^{M} \beta_m x_m^n$
- Least square loss function is given as $l(x, y) = \sum_{n=1}^{N} (y^n - \hat{y}^n)^2$
- Minimize the loss function $l(x, y)$ using Gradient Descent algorithm. Implement (learn-linregGD and minimize-GD algorithms given in the lecture slides). Choose i_{max} between 100 to 1000. Explain your choice [hint: the following plots might be useful in your choice.]
- You can choose three suitable values of step length $\alpha > 0$. For each value of step length perform the learning and record
  - In each iteration of the minimize-GD algorithm calculate $|f(x_{i-1}) - f(x_i)|$ and (when i_{max} is reached), plot it against iteration number i. Explain the graph.
  - In each iteration step also calculate RMSE on test set $RMSE = \sqrt{\frac{\sum_{q=1}^{T} (y_{test}^q - \hat{y}^q)^2}{T}}$ and plot it against iteration number i. Explain the graph.

Part B: (7 Points): Step Length for Gradient Descent

This task is based on Part A. You have to implement two algorithms *steplength − backtracking* and *steplength − bolddriver* given in the lecture slides. For each step length Algorithm

- In each iteration of the minimize-GD algorithm calculate $|f(x_{i-1}) - f(x_i)|$and plot it against iteration number i. Explain the graph.
- In each iteration step also calculate RMSE on test set $RMSE = \sqrt{\frac{\sum_{q=1}^{T} (y_{test}^q - \hat{y}^q)^2}{T}}$, plot it against iteration number i. Explain the graph.

Compare different step length algorithms Compare the RMSE graphs of $fixed - step - length$, $step - length - backtracking$ and $step - length - bolddriver$.

Explain your graph.

1.3 ANNEX

- You can use numpy
- 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.