Lab Course Machine Learning
Exercise Sheet 2

Prof. Dr. Dr. Lars Schmidt-Thieme, Shereen Elsayed
Information Systems and Machine Learning Lab
University of Hildesheim

November 12th, 2021
Submission on November 19th, 2021 at 12 noon, (on learnweb, course code 3116)

Instructions

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

1. You should submit a jupyter notebook detailing your solution.
2. Please explain your approach i.e. how you solved a given problem and present your results in form of graphs and tables.
3. Please submit your jupyter notebook to learnweb before the deadline. Please refrain from emailing the solutions except in case of emergencies.
4. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
5. Please refrain from plagiarism.

Exercise 1: Exploratory Analysis on Real-World Data using Pandas and Matplotlib (6 Points)

Like the subheading above suggests, in this task you are required to explore a real-world dataset from Rossman GmbH. The dataset was part of an online machine learning challenge where the task was to forecast sales of several stores. Please download and read about the attributes recorded for the dataset from here: `https://www.kaggle.com/c/rossmann-store-sales/data`. Specifically, you need the 'train.csv' and 'store.csv' files for this exercise sheet.

Part A: (Interesting stats)

- Find the store that has the maximum sale recorded. Print the store id, date and the sales on that day
- Find the store(s) that has/ve the least possible and maximum possible competition distance(s).
- What has been the maximum timeline a store has ran a "Promo" for? Which store was that, and what dates did the promotion covered?
- What is the difference in the mean of sales (across all stores) when offering a Promo and not?
- Are there any anomalies in the data as in where the store was "Open" but had no sales recorded? or vice versa?
- Which store type ('a','b' etc.) has had the most sales?

Part B: (Plotting)

- On a monthly basis how do the mean of sales vary (across all stores)? plot these sale
- On a daily basis how do the mean of sales vary (across all stores)? again, plot these sales.
- For the first store id, plot it's cumulative sales for the first year.
- Plot and comment on the following relationships:
  - customers(x-axis) vs. sales(y-axis)
  - competitiondistance(x-axis) vs. sales(y-axis)
- Plot an array of Pearson correlations between all features. Remember to do the merge operation between the dataframes store and train.
- For the first 10 stores (id'ed) draw boxplots of their sales
- From the above plot, which store has the highest median sales?

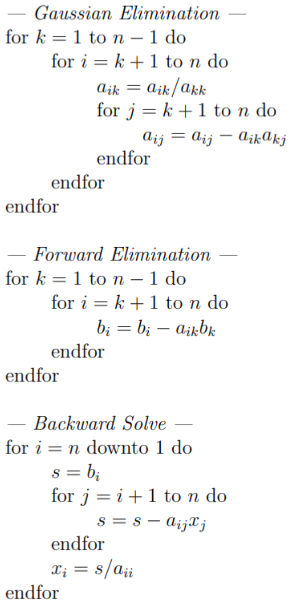
Exercise 2: Linear Regression (5+9 Points)

Part A: (Implementing Gaussian Elimination) We revisit the exercise sheet from last week and try to implement (multiple) Linear regression algorithm for 10 features.

1. Generate a simple data i.e. a matrix X with dimensions 100×10. Initialize it with normal distribution $\mu = 2$ and $\sigma = 0.01$
2. Generate a simple target vector i.e. a matrix Y with dimensions 100×1. Initialize it with random uniform distribution.
3. Implement linear regression algorithm and train it using matrix X to learn values of $\beta_{0:10}$. Let's denote β_0 as the parameter for the bias/intecept.
4. We already know an important aspect of this algorithm is solving the system of linear equations. For this exercise you are required to implement the algorithm given in Fig.1.
5. Implement the corresponding prediction algorithm and calculate the points for each training example in matrix X.
6. Plot the training points from matrix Y and predicted values $\hat{Y}$ in the form of scatter graph.
7. In the end use numpy.linalg.lstsq to learn $\beta_{0:10}$ and plot the predictions from these parameters.

Part B: Multiple Linear (Auto)Regression For this exercise, we shall try to forecast store sales from the Rossman dataset. Particularly, you are required to build multiple linear regression models to forecast sales for the next **42** days. Please follow the following steps:

1. Partition the provided train split as indicated in the following steps:
   a. Initialize X_{train}, Y_{train}, X_{test} and Y_{test}.
   b. For the first 1000 stores place all but their last **42** $Sales$ in X_{train}. Place the last **42** into Y_{train}. Generate X_{test} and Y_{test} accordingly from the remaining store ids. Also, please remove any stores that do not have sales recorded for **942** days.
   c. Print the shapes of these 4 data matrices. You should have shapes corresponding to dimensions of $(\#StoreIds, \#NumDays)$.



```
— Gaussian Elimination —
for $k = 1$ to $n - 1$ do
    for $i = k + 1$ to $n$ do
        $a_{ik} = a_{ik}/a_{kk}$
        for $j = k + 1$ to $n$ do
            $a_{ij} = a_{ij} - a_{ik}a_{kj}$
        endfor
    endfor
endfor

— Forward Elimination —
for $k = 1$ to $n - 1$ do
    for $i = k + 1$ to $n$ do
        $b_i = b_i - a_{ik}b_k$
    endfor
endfor

— Backward Solve —
for $i = n$ downto 1 do
    $s = b_i$
    for $j = i + 1$ to $n$ do
        $s = s - a_{ij}x_j$
    endfor
    $x_i = s/a_{ii}$
endfor
```

Figure 1: `Courtesy of http://www.math-cs.gordon.edu/courses/ma342/handouts/gauss.pdf. For more information you can go through the document.`

d. Iteratively build multiple linear regression models for column vectors of Y_{train}. You are allowed to use the *numpy* routines for calculating inverses, transposing of matrices and matrix multiplication.

e. Verify that you have learned $\beta_{0:900}$. Use these learned parameters to make predictions for each day ahead. In total 42 days.

f. Print one value for $RMSE$ and MAE each by aggregating the errors for all **42** days.

g. Compare this approach by reporting the error for the following baselines:

   g.1. Use the last recorded sales value of each store and repeat it for the next **42** days.

   g.2. Repeat the mean of sales for the sales horizon.

   g.3. For each of the 42 days ahead get their predictions as a mean of all sales recorded for that day of week in the past. For e.g. the prediction of Monday ahead should be the mean of all sales for this particular store on all previous Mondays.

h. Reason why or why not Linear Regression is a good choice for this task.