Lab Course: Distributed Data Analytics
Exercise Sheet 0

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

Submission deadline: Friday April 29, 23:59PM (on LearnWeb, course code: 3116)

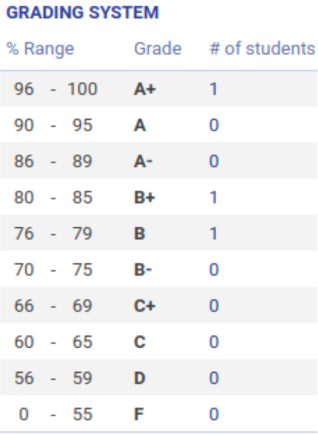
Instructions

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

1. You should submit 2 items 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. This is a warm up exercise but the points will count towards your final grade for this course.
5. Unless explicitly mentioned, you are not allowed to use scikit, sklearn or any other library for solving any part of the exercises. All implementations must be done by yourself.

Exercise 1: Pandas and Numpy (10 Points)

- **Matrix Multiplication**: Create a numpy matrix A of dimensions $n \times m$, where $n = 100$ and $m = 20$. Initialize Matrix A with random values. Create a numpy vector v of dimensions $m \times 1$. Initialize the vector v with values from a normal distribution using $\mu = 2$ and $\sigma = 0.01$. Perform the following operations:
    1. Iteratively multiply (element-wise) each row of the matrix A with vector v and sum the result of each iteration in another vector c. This operation needs to be done with for-loops, not numpy built-in operations.
    2. Find mean and standard deviation of the new vector c.
    3. Plot the histogram of vector c using 5 bins.
- **Grading Program**: This task puts you in the position that I end up at the end of every semester. Which is, grading your work and issuing the grades. In this task you are required to use the 'Grades.csv' File that has been provided on learnweb.
- Read the data from the csv.
- For each student,
    - Compute the sum for all subjects for each student.
    - Compute the average of the point for each student. (total points are 500).
    - Compute the standard deviation of point for each student.
    - Plot the average points for all the students (in one figure).
    - For each student assign a grade based on the following rubric.
    - Plot the histogram of the final grades.



GRADING SYSTEM

% Range	Grade	# of students
96 - 100	A+	1
90 - 95	A	0
86 - 89	A-	0
80 - 85	B+	1
76 - 79	B	1
70 - 75	B-	0
66 - 69	C+	0
60 - 65	C	0
56 - 59	D	0
0 - 55	F	0

Figure 1: Grading Rubric

Exercise 2: Linear Regression through exact form. (10 Points)

In this exercise, you will implement linear regression that was introduced in the introduction Machine Learning Lecture. The method we are implementing here today is for a very basic univariate linear regression.

- Generate 3 sets of simple data, each consisting of a matrix A with dimensions 100×2. Initialize the sets of data with normal distribution $\mu = 2$ and $\sigma = [0.01, 0.1, 1]$ so that each dataset has a different σ. You may assume that the first column of A represents the predictor data (X), whereas the second column of matrix A represents the target data (Y).
- Implement LEARN-SIMPLE-LINREG algorithm and train it using matrix A to learn values of β_0 and β_1
- Implement PREDICT-SIMPLE-LINREG and calculate the points for each training example in matrix A.
- Plot the training data (use plt.scatter) and your predicted line (use plt.plot).
- Put β_0 to zero and rerun the program to generate the predicted line. Comment on the change you see for the varying values of σ
- Put β_1 to zero and rerun the program to generate the predicted line. Comment on the change you see for the varying values of σ
- Use numpy.linalg lstsq to replace step 2 for learning values of β_0 and β_1. Explain the difference between your values and the values from numpy.linalg lstsq.