Introduction to Machine Learning (IML)
BPINFOR-113
Take-Home Assignment 1 (THA1)
Grading scheme: 20 points (0-20)
Weight for final grade: 10%

Responsible: Decebal Mocanu, Christopher Gadzinski - SL, Patrick Keller - UL

Release date: 6 October 2023.
Deadline: 30 October 2023, $23:59$.

Preamble

Please submit your solutions (the developed code and a report in PDF format) through Moodle. State clearly in the report the group number and the group members.

The solutions for the Take-Home Assignment 1 can be submitted just by a group. Individual submissions are not allowed. Each group shall have three members.

A. Supervised Learning (SL) - Regression (8 points)

Problem

In this exercise (A) you are requested to perform linear and polynomial regression on the data collected from a Combined Cycle Power Plant. The data can be found in Moodle in the file "Data Take Home Assignment 1 Exercise A.xlsx“. It contains two columns. The first column represents the *Temperature* (the input data - X) on the Celsius scale and the second column represents the *Net hourly electrical energy output* (the output data - Y) in MW. Each row represents a data point.

Tasks

- Please solve each of the following subproblems and give clear explanations for your solutions.
- For this exercise (A), you are requested to develop the code for Least Squares, Linear Regression and Polynomial Regression optimised with Gradient Descent from scratch. That is: do not use any libraries such as Scikit-Learn to implement these machine learning models for regression, the equations that you will report in the pdf report to calculate the derivative of the loss functions, the gradients, the update rules, and so on, shall be clearly identifiable in the developed code.

A.I. Linear regression (5 points)

1. (0.5 points) *Data acquisition.* Each group has to perform linear regression on 20 data points (not on all data points from the file). Assuming that your group number is n then the data points corresponding to your group starts with row number $(n-1)*20+2$ and ends with the row number $n*20+1$ (e.g. Group 1 has the data from row 2 until row 21, Group 2 has the data from row 22 until row 41, and so on). Present your data in a table in the report.

2. (0.5 points) *Data transformation.* Perform a transformation of your acquired data. Please feel free to use any type of transformation you prefer (e.g. Min-Max normalization, Z-score standardization). Report the formula that you used for the transformation and the transformed data in the report. Justify your choice.

3. (0.5 point) *Least Squares.* Perform linear regression on your data using least squares (closed form solution). Print the obtained regression parameters. Plot the regression line and the data points.

4. (1 point) *Linear Regression with Gradient Descent - cost function.* Choose a suitable cost function to perform linear regression with gradient descent. Motivate your choice. Perform the partial derivatives of the cost function with respect to each of the regression parameters. Report also the update rules.

5. (0.5 points) *Linear Regression with Gradient Descent - first iteration.* Start from a random choice of the regression parameters (print the choice in the report) and perform the first iteration of gradient descent on your data. Report the chosen learning rate and the computed values. Plot the regression line and the data points.

6. (0.5 points) *Linear Regression with Gradient Descent - second iteration.* Perform the second iteration of gradient descent. Report the computed values. Plot the regression line and the data points.

7. (0.5 points) *Linear Regression with Gradient Descent - third iteration.* Perform the third iteration of gradient descent. Report the computed values. Plot the regression line and the data points.

8. (0.5 points) *Linear Regression with Gradient Descent - last iteration.* Continue performing gradient descent for an arbitrary number of iterations until the regression line fits well the data. Plot the values of the cost function for all iterations. Report the regression parameters and plot the regression line that you obtained after the last iteration.

9. (0.5 points) *Discussion.* Compare the performance of Least Squares and Linear Regression optimised with Gradient Descent. Discuss the similarities and the differences between them.

A.II. Polynomial regression (3 points)

- *Discussion.* Using the data from subproblem numbers A.I.1 or A.I.2 (this is your choice, but please justify it) please derive, implement, and perform polynomial regression using the following details:
- *Regression model:* $h_\theta(\mathbf{x}) = \theta_2 x^2 + \theta_1 x + \theta_0$, where $\theta_2, \theta_1, \theta_0,$ are the model parameters
- *Cost function:* $J(\theta) = \frac{1}{4n}\sum_{i=1}^{n}(y^{(i)} - h_\theta(\mathbf{x}^{(i)}))^4$, where i iterates over all data points n, $\mathbf{x}^{(i)}$ represents the input of a data point (i), $\mathbf{y}^{(i)}$ represents the true output of a data point (i)
- *Optimization method:* Gradient descent

The report shall contain:

1. (0.5 points) Partial derivatives of the cost function with respect to each of the model parameters.

2. (0.5 points) The update rule for each parameter

3. (1.5 points) Plots to reflect how during the gradient descent iterations the model fits better and better the data. Stop the iterations when the model fits well the data. Please report also the initial values of the model parameters and the learning rate.

4. (0.5 points) A discussion with the differences and similarities between linear and polynomial regression.

B. Supervised Learning (SL) - Classification (5 points)

Problem

In this exercise (B) you are requested to create two artificially generated datasets (as detailed below) and to use the scikit-learn library to visualize the decision boundaries of k-NN, Naive Bayes, Decision Trees, and Random Forests. Even if it was not discussed during the lectures, please feel free to add also a Support Vector Machine and an Artificial Neural Network classifier.

Tasks

- Please solve each of the following subproblems and give clear explanations for your solutions.
- For this exercise (B), you do not have to implement the machine learning models from scratch. Some hints: try to use the scikit-learn library; this tutorial https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html can be used as a guideline to solve this exercise.

B.I. Classification models comparison (4 points)

1. (0.5 points) Dataset 1 creation. The dataset shall represent a binary classification problem and shall have as input two continuous real number attributes. The dataset shall have 40 data points belonging to one class and 40 data points belonging to the other class. Discuss your approach and report the data points obtained in the report.
2. (0.5 points) Dataset 2 creation (Dataset 1 + outliers). Add randomly to Dataset 1, four outliers for each class. Discuss your approach and report the data points, including the outliers in the report.
3. (0.5 points) Split Dataset 1 and Dataset 2 in training and testing data. Justify your choices for splitting the data and report the split data.
4. (1 point) Train on Dataset 1 and Dataset 2, the following classification models: k-NN, Naive Bayes, Decision Trees, and Random Forests. Report the classification accuracy obtained on the training and on the testing data. Report also the confusion matrix. Discuss the obtained results.
5. (1 point) Taking inspiration from the tutorial mentioned above, visualise the decision boundaries of the various classification models used.
6. (0.5 points) Discuss the similarities and the dissimilarities between the various classification models. Include in the discussion also their decision boundaries.

C. Unsupervised Learning (UL) (8 points)

Make sure you use fixed random seeds - If we are not able to run and reproduce your results you may lose points. If you put multiple answers into the same .py file, add a comment section that clearly denotes which subtask the code belongs to, indicate the line-number where the answer starts in the report.

C.I. Iris clustering with K-means (4.5 points)

1. (0.5 points) Study how the K-means algorithm works and describe it in 3-5 short sentences in your own words in your report.
2. (0.5 point) Run k-means as provided by the scikit-learn library on the iris dataset used in the lecture before. Implement your solution in kmeans_iris.py Plot the data columns pairwise as scatter plots, colored by the different clusters determined by k-Means.
3. (1.0 point) Using the data labels, compute and report the accuracy and the confusion matrix for the produced clusters. Motivate based on these results whether the trained model is suitable for this problem.

4. (0.5 point) Compute the average silhouette score for each cluster and describe what this metric means.

5. (1.0 point) unknown_species.csv contains 6 iris flowers whose species are unknown yet. Use your trained k-means model to predict which cluster these new flowers belong to. In the report provide a list that assigns the iris species determined by K-means to each id of the unknown flowers. Also, compute (in code) and report the silhouette score for each individual flower and explain the meaning of the silhouette score.

6. (1.0 point) Implement (naive) k-means from scratch in the file my_kmeans.py and evaluate your implementation on the iris dataset. Evaluate how your implementation performs compared to the one provided by scikit-learn.

C.II. Clustering of three-dimensional shape dataset using DBScan (3.5 points)

1. (0.5 points) Study the DBScan algorithm and describe how it works in 3-5 sentences.

2. (0.5 points) Download the ulu.csv dataset from Moodle and run the DBScan algorithm from the scikit-learn library on it. How many clusters does it generate with default parameters? Put your code into ulu_dbscan.py.

3. (0.5 points) Plot scatter plots per pair of features of the produced clusters using different colors. What shapes can you recognize?

4. (0.5 points) Run k-means on the dataset and explain why the clusters do not make sense using this algorithm.

5. (0.5 points) Compute the silhouette score for the results of DBScan and K-means. Why is the silhouette score worse for DBScan even though the clusters make more sense?

6. (0.5 points) Run a PCA dimensionality reduction on the dataset and plot the results colored by cluster as determined by DBScan.

7. (0.5 points) What transformation does the PCA results reflect compared to the original dataset. (Hint: plot the dataset in 3D)

Success!