Assignment 1

Innopolis University

Introduction to Machine Learning, Fall 2022 - Bachelors

General Instructions

In this assignment, you will solve an interesting machine learning task where you will be required to implement and compare various regression and classification techniques. The task is divided into two parts: practical task and theoretical part. In the practical task, you will solve a classification task where you are required to implement regression techniques in the initial phase, the details will be explained the Section 1; the Section 2 is about the theoretical part, where you have to answer the questions related to the work that you do in the practical task.

You are required to submit your solutions via Moodle as a single zip file (only ZIP format is accepted; neither RAR nor any other archive format). The zip archive should contain a single *.ipynb* file, and a single PDF for the theoretical section. Please, put your name and email at innopolis.university as the first line in the notebook and in the report document as well.

Deadline: 23:00 (MSK), 2nd of October, 2022.

Do not just copy and paste solutions from the Internet. Plagiarism will be checked and measures will be taken accordingly. You are allowed to collaborate on general ideas with other students as well as consult books and Internet resources. However, be sure to credit the sources you use and type all the code, documentation by yourself.

Dataset

You will be working on the dataset **"a1_dataset.csv"** throughout the practical task. The following is the information about the dataset:

- there are 925 rows in the dataset
- the *var1, var2, var3, ... var7* columns are features, while the *target* column is the label that should be predicted
- the *var3, var6* columns are categorical values;
- the values of the *var7* column are in datetime format;
- the *var1, var2, var4, var5* columns are in numerical format
- the *var4* column is missing around 600 data points

The Figure 1 shows how the dataset looks like. You can see that the feature names (*var1, var2 etc.*) do not make any sense. This is a common practice that you might encounter such datasets which lack column names (in reality, there might be various reasons - security reasons where companies do not want to expose information due to the confidentiality policy, or companies do not want interns to be biased towards a dataset by its column names etc.)

target	var1	var2	var3	var4	var5	var6	var7
0	509.180	417.681	Micronesia	138.000	393.000	no	2019-07-20 13:21:37
0	446.060	666.182	Dominica	81.000	352.050	yes	2019-04-04 21:30:46
1	235.500	398.097	Isle of Man	90.000	339.000	no	2019-03-03 02:59:37
0	306.020	518.163	Turkmenistan	102.000	439.250	yes	2019-03-19 08:00:58
0	453.080	600.156	Cameroon	105.000	422.950	no	2019-03-18 13:22:35
1	211.720	506.716	Liechtenstein	111.000	310.600	no	2019-03-18 13:00:12
0	401.420	627.294	French Guiana	78.000	390.050	no	2019-03-28 02:29:19
0	498.900	525.207	Barbados	129.000	408.750	yes	2019-06-07 05:41:16

Figure 1: A1 dataset preview

1 Practical task (80 %)

The practical task is also divided into Preprocessing and Training parts that you should follow in order to accomplish the task.

1.1 Preprocessing (50 %)

Encoding categorical values

The *var3* and *var6* columns of the dataset have categorical values which should be encoded. Implement the categorical encoding techniques that you learned during the labs for this task. Make sure that you try the Ordinal or One-hot encoding techniques. You can also try both of them and compare their efficiency, then select the best performing encoding technique. You may also try other encoding techniques, but make sure that you try at least one of the encoding techniques mentioned above.

The *var7* contains values in date-time format. But you are free whether to drop this column or encode and use it in the further classification model. So, you are not required on handling the column *var7*. However, if you handle this column as well, then you get extra points.

Data imputation

As stated earlier, the *var4* column is missing around 600 values. This issue is also common in practice (maybe dataset is corrupted, or not enough data on certain columns etc.), and you should be able to address. There are several statistical imputation techniques for missing values, such as filling with the column mean or median or even with some constant value etc.

But in this task, you are required to implement regression techniques (since the values of the column is continuous) to find the missing values of the *var4* column. So, in the regression task, you have to make the *var4* column as target, while you utilize the rest of the columns as predictors.

You are free to choose any regression model, but make sure that you try the following regression models:

- Linear regression
- Polynomial regression (with at least 3 degrees, for example within $[2, 4]$ range etc.)

Finally, select the most optimal regression model and fill in the missing values.

Implementing the PCA technique

Visualisation is an important phase in data pre-processing which can reveal hidden patterns or crucial information about features, as well as relationship among them. So you have to properly visualise the data. But, as you can see that there are more than three features in the dataset which makes it hard to effectively visualise the data without reducing the dimensionality.

Here, you have to implement the PCA technique for reducing the dimensionality. During the labs, you learned how to implement the PCA using sklearn. But in this assignment, you are required to implement it from scratch following the steps described in the lectures. You should show different scatter plots that visualises the data.

1.2 Training (50 %)

Once you have preprocessed the dataset, the data is ready for the training process. In the training process, you have to implement several classification techniques. You treat the *target* column as target and try to predict its values.

Implement the following models for classification and compare their performances to select the best performing model:

- Logistic Regression
- KNN (try the $n\_neighbors$ hyperparameters between $[1, 10]$ range)
- Naive Bayes

When implementing the ML algorithms, make sure to fine-tune the hyper-parameters properly as you learned during the classes by evaluating the ML models on the cross validation set that you obtain by implementing the K-Fold Cross-Validation technique. You should set $k = 3$ for the K-Fold Cross-Validation technique and calculate the average performance of the model on cross-validation set.

In the Preprocessing part, you were asked to implement the PCA for data visualisation. In the training task, you are also required to implement the PCA, but this time for reducing the dimensions to test if model performs better in this case. Overall, first you have to train all the above mentioned classification models with existing (pre-processed by you) features, then secondly, try the above classification models with reduced dimensions by applying the PCA.

2 Theoretical part (20 %)

Write a report about the work that you did in the Practical task section. Make sure that you address the following questions in your report.

2.1 Regarding the Preprocessing (20 %)

- Which regression model was the most effective for the missing values, and why? (50 %)
- What encoding technique did you use for encoding the categorical features, and why? (50 %)

2.2 Regarding the training process (80 %)

- Which classification model performed best, and why? (30 %)
- What were the most critical features with regards to the classification, and why? (20 %)
- What features might be redundant or are not useful, and why? (20 %)
- Did the dimensionality reduction by the PCA improve the model performance, and why? (20 %)

Additional research: (a) what is a multi-label learning problem? (b) suggest an example in which you can transform the given problem into a multi-label problem? Will the models work as it is in that case, or would some changes be required? (10 %)

Notes

- Cheating is a serious academic offense and will be strictly treated for all parties involved. So delivering nothing is always better than delivering a copy.
- Late assignments will not be accepted and will receive **ZERO** mark.
- Code cleanness and style are assessed. So maybe you want to take a look at our references: Link 1 and Link 2.
- Organize your notebook appropriately. Divide it into sections and cells with clear titles for each task and sub-task.