Birzeit University

Department of Electrical & Computer Engineering

First Semester, 2022/2023

ENCS3130 Linux Laboratory

Shell Scripting Project – Data Preprocessing

In this project, you are required to write a shell script that encode and scale features of a dataset written in a text file.

Dataset

The dataset is a set of features arranged as columns that are separated by a delimiter. In this project, the delimiter is the semicolon ";" as shown in the figure below. The first row of the dataset is the header which presents the names of the features. The features in the dataset can be of two types; (1) numeric features of type integer such as age, height, and weight, and (2) categorical features such as gender, active, smoke, and governorate.

```
id;age;gender;height;weight;active;smoke;governorate;
1;30;male;170;88;no;yes;ramallah;
2;25;female;160;65;no;no;ramallah;
3;28;male;165;72;yes;yes;nablus;
4;44;male;188;90;no;no;jerusalem;
5;60;female;166;70;no;no;jerusalem;
```

Figure 1: A snapshot of a dataset in a text file.

Dataset is Clean

In this project, we will assume that the dataset has been cleaned. I.e., all rows in the dataset contain values for all features with the correct data type and there are no missing values.

Encode Features

For categorical features, there are two types of encoding; label encoding and one-hot encoding. The description of each of these encodings is as follows:

1) Label encoding replaces categorical data with integer codes. As an example, the label encoding of the governorate features in the dataset of figure 1 will replace ramallah with 0, nablus with 1, and jerusalem with 2 as shown in figure 2.

```
id;age;gender;height;weight;active;smoke;governorate;
1;30;male;170;88;no;yes;0;
2;25;female;;65;no;no;0;
3;28;male;165;72;yes;yes;1;
4;44;male;188;90;;no;2;
5;60;female;166;70;no;no;2;
```

Figure 2: label encoding of the governorate feature

2) One-hot encoding splits the categorical feature column into multiple columns and each sample is encoded by 0 or 1. As an example, the governorate feature in the dataset of figure 1 is replaced with three column features; ramallah, nablus, and jerusalem. Also, the code of these new features for the sample with id=1 who lives in Ramallah is 1;0;0 as shown in figure 3.

```
id;age;gender;height;weight;active;smoke;ramallah;nablus;jerusalem;
1;30;male;170;88;no;yes;1;0;0;
2;25;female;;65;no;no;1;0;0;
3;28;male;165;72;yes;yes;0;1;0;
4;44;male;188;90;;no;0;0;1;
5;60;female;166;70;no;no;0;0;1;
```

Figure 3: one-hot encoding of the governorate feature

Feature Scaling

In this project, we will use the MinMax Scaling technique. The equation for the MinMax Scaling of the feature is given as:

$$x_{i,scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Where x_i is the value of the feature vector x, x_{min} and x_{max} are the minimum and maximum values of x, and $x_{i,scaled}$ is the MinMax scaled version of feature value x_i. As an example, for the feature vector $x = [1, 2, 3, 2, 2]$, the MinMax scaled feature $x_{scaled} = [0.0, 0.5, 1.0, 0.5, 0.5]$.

Procedure:

1. The program should print on the screen the main menu and ask the user to select an option

```
r) read a dataset from a file
p) print the names of the features
l) encode a feature using label encoding
o) encode a feature using one-hot encoding
m) apply MinMax scalling
s) save the processed dataset
e) exit
```

Figure 4: the main menu of the Data Preprocessing project

2. If the user enters '**r**':
   a. The program should print on the screen "Please input the name of the dataset file".
   b. The program should verify that the file exists, otherwise a message should be printed on the screen "file does not exist" and then return to the main menu.
   c. The program should then check the format of the data in the dataset file according to the description above. In case of any format problems, the program should print on the screen "The format of the data in the dataset file is wrong" and then return to the main menu.
   d. If the person selects any option other than 'r' or 'e' before the format of the data in the dataset file is verified correctly, the program should print on the screen "You must first read a dataset from a file" and then return to the main menu.
3. If the user enters '**p**', the program should print on the screen the names of all features of the dataset file and then return to the main menu.
4. If the user enters '**l**':
   a. The program should ask for the name of the feature to be encoded using label encoding by printing on the screen "Please input the name of the categorical feature for label encoding".
   b. The program should verify that the entered name of the categorical feature exists in the dataset, otherwise prints on screen "The name of categorical feature is wrong" and then return to the main menu.

c. If the entered name of the categorical feature exists, the program should print on the screen the distinct values of the categorical feature and the code of each value. And also, to encode the categorical feature in the dataset using label encoding as described above and then return to the main menu.

5. If the user enters '**o**':
   a. The program should ask for the name of the feature to be encoded using one-hot encoding by printing on the screen "Please input the name of the categorical feature for one-hot encoding".
   b. The program should verify that the entered name of the categorical feature exists in the dataset, otherwise the program should print on screen "The name of the categorical feature is wrong" and then return to the main menu.
   c. If the entered name of the categorical feature exists, the program should then print on the screen the distinct values of the categorical feature. And also, to encode the categorical feature in the dataset using one-hot encoding as described above and then return to the main menu.

6. If the user enters '**m**':
   a. The program should ask for the name of the feature to be scaled using MinMax scaling by printing on the screen "Please input the name of the feature to be scaled".
   b. If the entered feature is a categorical feature, the program should verify that this feature is encoded, otherwise, the program should print on screen "this feature is categorical feature and must be encoded first" and then return to the main menu.
   c. If the feature is numeric or encoded categorical feature, the program should print on the screen the minimum and maximum values of the feature and apply the MinMax scaling to the feature vector and then return to main menu.

7. If the user enters '**s**':
   a. The program should print on the screen "Please input the name of the file to save the processed dataset".
   b. The program should save the processed dataset into the entered filename and then return to the main menu.

8. If the user enters '**e**':
   a. The program should check if the processed dataset is saved using option "s". if not, the program should print on the screen "The processed dataset is not saved. Are you sure you want to exist". If the person inputs "yes", the program ends. Otherwise, the program should return to main menu.
   b. However, if the dataset is saved, the program should print on the screen "Are you sure you want to exist". If the person inputs "yes", the program ends. Otherwise, the program should return to the main menu.

Submission:

Please submit the following:

1. Shell script program
2. Report: the report must include:
   a. The code, idea, and a screen shot of each option with its variations. For example: for the option "e" you need to add code + description + screen shot of the output for cases (a) and (b) under (e).

b. At least 2 dataset test examples.

Notes:

- Write the code for the shell script to satisfy the requirements described above and name the script as DatasetPreprocessing.sh.
- Make sure your code is clean and well-indented; variables have meaningful names, etc.
- Make sure your script has enough comments inserted to add clarity.
- Work in groups of at most two students
- Deadline: Thursday, 29 December 2022 at 11:59 pm. Please submit your project (code + report) through Ritaj as a reply to this message.
- This project is per group effort: instances of cheating will result in you failing the lab.