ADAT 5340: Midterm – Important Notes

Thuan L Nguyen, PhD

1. Overview

To give all of you some help while you work on the last section of the midterm, I will discuss some important details of coding for this part of the midterm.

2. Dataset: Preprocessing

In the dataset **loan_approval.csv**, there is one attribute, **Dependents** (number of dependents), that has numeric values such as 0, 1, 2, … The problem with this attribute is that its values are not consistent: Most are numeric (integers). However, several are strings (3+). Such inconsistency would cause big problems for the process. It should be included in the cleaning process.

There are many ways to fix it. The simplest option is to change the string "3+" to the integer 3 that means **"3 or more."** It means that other values (0, 1 ,2) would mean the exact number of dependents. However, the value of 3 would mean **"three or more dependents."**

How to change the string "3+ to the integer '3'?
You can write some code to do it.
You still can do manually (as removing the records with missing values).

The student needs to address this problem. Otherwise, some error would occur down the road.

3. PART IV: Coding

First, for all the coding (not only PART IV), if you want to suppress the future warnings (so many) in the results when you run the code, you can add the following lines of code at the top:

```
//---------------
import warnings
warnings.filterwarnings("ignore")
//---------------
```

Second, after cleaning the dataset, when you run the code of PART IV (loan_approval.csv) to build and train the model, some of you might see the following error or similar to it:
"… cannot convert to float: Male …"

I mentioned "**some of you**" because it does not occur to everyone, maybe due to something happening in the embedded sw of your desktop or laptop.

How to solve this problem: You should perform the integer encoding for every categorical attribute in the dataset. Here is I give you the sample code to do it:

Let's look back at the code in the video lectures of the logistic regression.

When the dataset is ready for the ML tasks, it will be split into two subsets: One for the predictors, and one for the labels, i.e., outcomes.

Here is the original code in the video lectures:

```
//---------------
# Store dataframe values into a numpy array
array = df.values

# separate array into input and output components by slicing
# For X (input)[:, 1:5] --> all the rows, columns from 1 - 4 (5 - 1)
X = array[:,1:5]

# For Y (input)[:, 5] --> all the rows, column 5
Y = array[:,5]
//---------------
```

To solve the above issue, you need to add the following lines of code in a new cell above the cell of the displayed code to integer-encode the categorical attribute "Species":

```
//----------------
df.Species = pd.Categorical(df.Species)
df['Species'] = df.Species.cat.codes
//----------------
```

By adding these line of code, all the categorical values of the categorical attribute are converted into numeric values. Now the values of the last attribute have been encoded into integers and the data type is changed (from Pandas view)

$ df.dtypes

```
Id                 int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species             int8
dtype: object
```

The problem is still not yet solved for some of the users because **Numpy still does see these values as strings, not yet as integers**. You still need to **add one more line of code** to completely solve the issue:

```
//--------------
# Store dataframe values into a numpy array
array = df.values

# separate array into input and output components by slicing
# For X (input)[:, 1:5] --> all the rows, columns from 1 - 4 (5 - 1)
X = array[:,1:5]

# For Y (input)[:, 5] --> all the rows, column 5
Y = array[:,5]
Y = Y.astype('int')
//---------------
```

Now you can solve the issue entirely.

Please be reminded that I’ve shown you the sample code using the Iris project in the video lectures so that you can learn from them and apply similar code into the midterm/PART IV. You don’t have to do anything with Assignment 3.

Good luck with the midterm.

Dr. Nguyen