ADVANCED PROGRAMMING
MASTER IN STATISTICS FOR DATA SCIENCE. 2019-20
Basic Python Programming: Categorical Mean Encoding

1.0 POINTS

Introduction

In this assignment, we are going to program a simple version of "Categorical Encoding" using **base** Python. That means that more advanced data types such as numpy matrices or Pandas dataframes are not going to be used here.

Categorical mean encoding is a useful technique (but not widely known), to deal with categorical variables that contain many different values. For these variables, dummy variable encoding does not work well. Basically, the idea is to replace each categorical value by the average of the corresponding values of the response variable. Obviously, there are Python libraries for carrying out different versions of categorical mean encoding, but in this case we are going to program it for the sake of practicing base Python programming.

In principle, a data matrix contains several features/predictors/attributes and one response variable (or target). For simplicity's sake, we are given only one attribute and the response variable, which is a real number.

The dataset we are going to use is the "Auto MPG dataset", whose goal is to predict a numerical variable (Miles per Galon) in terms of several attributes such as "number of cylinders", "weight", etc. One of the attributes is the car model year, with values from 70 to 82. Although it is an integer, it can be considered a categorical variable.

The aim of the assignment is, using base Python, program categorical mean encoding by following the next steps.

First, we load the dataset and obtain the response variable and the "car model year" (column number 6). The best way to load complete datasets into Python is using Pandas dataframes, but given that we haven't studied them yet, I'll provide the code here:

```
import pandas as pd
cars = pd.read_csv('auto-mpg.data', delim_whitespace=True,header=None)
cars.describe()

# The response variable is mpg (miles per gallon)
# Attribute 6 is the car model year: 70, 71, 72, ..., 82

modelyear = cars.iloc[:,6].tolist()
mpg = cars.iloc[:,0].tolist()
```

Second, use a dictionary to associate each categorical value (in *modelyear*) with a list that contains all the response variable values for that categorical value.

Third, use the *reduce* function to compute the average of the response variable values (for each categorical value).

Fourth, create a **copy** of the response variable vector, and in that copy, replace each categorical value by the corresponding response-variable average for that categorical value.

What to hand in:

A text file containing the Python code, and another file containing the values of the encoded categorical variable.

More information about categorical encoding:
https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02