Auxiliary Feature Learning for Small Dataset Regularization (January 2019)

Jordà. Jaume, *Data Science and Decision Making, University of Essex*

***Abstract* – An autoencoder is a neuronal network able to generate new data.**
The autoencoder can learn by a few data and then, with different machine learning methods like lineal regression, classification, etc the neural network can take some data of the same type and classify it, obviously, with some error.
With three different small datasets we will understand how an autoencoder works.

I.INTRODUCTION

The autoencoder consists of two parts, the encoder and the decoder, the encoder compress the input data in a latent variable space and the decoder try to rebuild the input based in previous information that it recollects.

The purpose of this paper is the selection of the datasets to train the autoencoder, it is done based on the number of attributes that every dataset has, because with fewer attributes we will understand it better. Also, we need to use small datasets (number of instances) or if it is not possible, we must choose a little sample of the dataset.

The datasets are provided by *UCI Machine Learning Repository* and the choose datasets are *flags of the* world, *wines*, and *zoo*.

II.BACKGROUND

To start working on these datasets we must know the previous works that the creators of the datasets did with it.

The *flags* and *zoo* datasets were used only in the document that Forsyth's, the creator of the datasets, used to explain what is BEAGLE and how he program it called "BEAGLE User Notes" and we can find it here.

The *wines* dataset is created by Forina, M from the Institute of Pharmaceutical and Food Analysis and Technologies, and it was used in 1992 by S. Aeberhard, D. and O. de Vel of University of North Queensland to compare different classifiers in high dimensional settings and for a paper called "The Classification performance of RDA".

III.METHODOLOGY

Before start working it is necessary how are the datasets.

The *flags* dataset consists on 194 instances of 30 attributes. The attributes are the name, the landmass, the graphic quadrant zone, the area, the population, the language, the religion, the bars in the flag, the stripes in the flag, the number of different colours, (red, green, blue, gold/yellow, white, black, orange) as a boolean if is present or not in the flag, the predominant colour in the flag, the number of circles, the number of upright crosses, the number of diagonal crosses, the number of quartered sections, the number of sun or star symbols, is crescent moon present, is triangle present, if there are inanimate symbology (e.g., a boat), if there are animate image like eagle, if there are letters or writing, if there is colour in the top-left corner, and finally, if there is colour in the bottom-left. Is possible to see that this dataset is a little bit large but before start working we have to select a sample.

The *wines* dataset is distributed in three classes, the class 1 have 59 instances, the class 2 have 71 and the class 3 have 48, and every instance have 13 attributes. The attributes are the alcohol, the malic acid, the ash, the alkalinity of the ash, the magnesium, the total phenols, the flavonoids, the nonflavanoid phenols, the proanthocyanins, the color intensisty, the hue, the OD280/OD315 of diluted wines, and the proline.

The *zoo* dataset have 101 instances of 18 attributes, the animal name, 15 boolean attributes, and two numeric values. The boolean attributes make sense to if the animal is/have/do something like hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic and catsize, and the numeric values are the number of legs (set of values: {0, 2, 4, 5, 6, 8}) or the type (integer in range [1, 7].

IV.EXPERIMENTS

The experiment I did with the datasets is simply charge the data to the program and print it in the screen. The aim to do that is to be sure that the choosed datasets will be useful to learn how an autoencoder works.

In the images below we can se the piece of code that allows us to do it and how is shown in the screen with the *flag* dataset to show the example.

dataReader.py

```
import numpy as np
#import panda as pd
import matplotlib.pyplot as plt

# Open the data file in reading mode
f = open("flag.data", "r+")

# Save the information in the data variable
data = f.read()

# Show de data on the screen
print(data)

f.close # Close the data file
```

V.DISCUSION

Once the datasets are selected it is important to select how much data is necessary to train the autoencoder and how much is for testing.

In this case the 80% will be used as training data and the remaining 20% for tests, that is, see if the autoencoder is well trained because it is about unsupervised training.

VI.CONCLUSION

Finally we can see that the datasets are enough for our proposes, so in the next paper we can see if our proposal it is good or not.

VII.NOTES

To see the github repository it is necessary to use this link:

https://github.com/jj18367/ce888_assignment1.git