Some Analysis of the Data

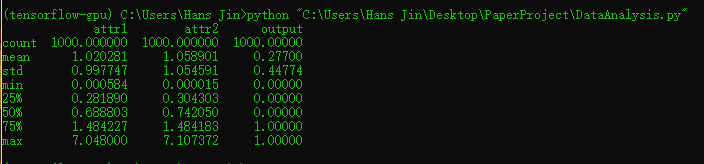
(2019/3/25)

In computer programming, **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

In the very beginning, I need to separate the original data into 3 columns, attr1, attr2, and output. And then put the separate data into data.csv file, so that I can use python to analyze them.

***df = pandas.read\_csv(datapath)***

***print(df.describe())***

This method can get the mean value, standard deviation, min and max of the 3 columns. The result is shown below.

***print(df[df.output==0])***

This can print out the row whose output is 0.

And also, it shows that there are 723 lines, which means that the output has 723 0s.

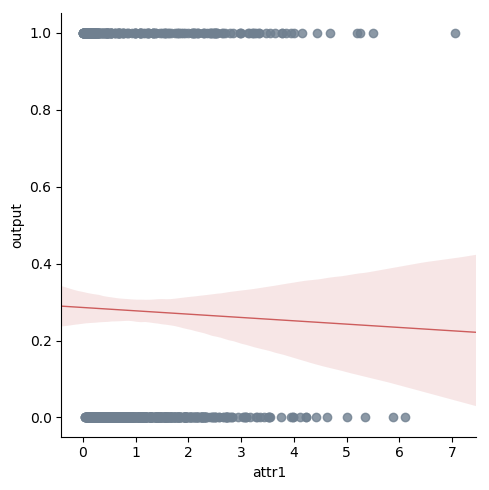
***print(df[df.output==1])***

This is for the row whose output is 1.

***print(df.sort\_values(by=‘attr1’))***

This can put attr1 in up-order and then print out.

***seaborn.lmplot("attr1","output",df, scatter\_kws={"marker": ".", "color": "slategray"}, line\_kws={"linewidth": 1, "color": “indianred"}).savefig('picture2')***

This can generate a picture which shows the relation between attr1 and output. But until now, I can not find any useful information from the picture.

(2019/3/26)

***sns.lmplot("attr3","output",df, scatter\_kws={"marker": ".", "color": "slategray"}, line\_kws={"linewidth": 1, "color": "indianred"}).savefig('OutputToAbs')***

This can generate a picture that shows the relation between attr3 and output, where attr3 is the absolute value of the difference between attr1 and attr2. The picture is shown below.

From the picture, we can see that the curvilinear regression looks like a straight line, which is similar to :

output = 1/4\*attr3

In my opinion, when 1/4\*attr3 is less than 0.5, the real output will be 0. From the attr3 axis, it seems that the result will hold for most attr3.

