STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

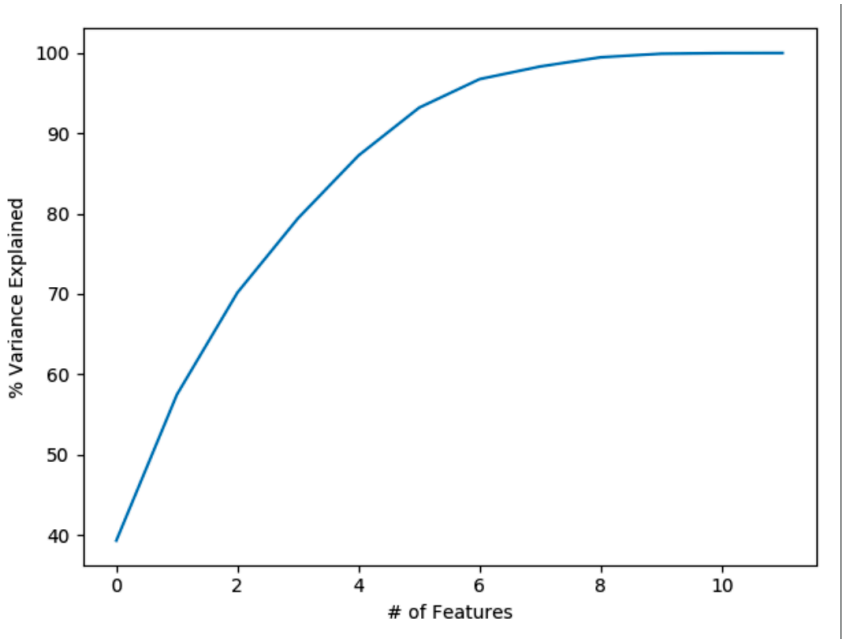
Signature: Haodong Zhao Date: Apr 8th. 2019

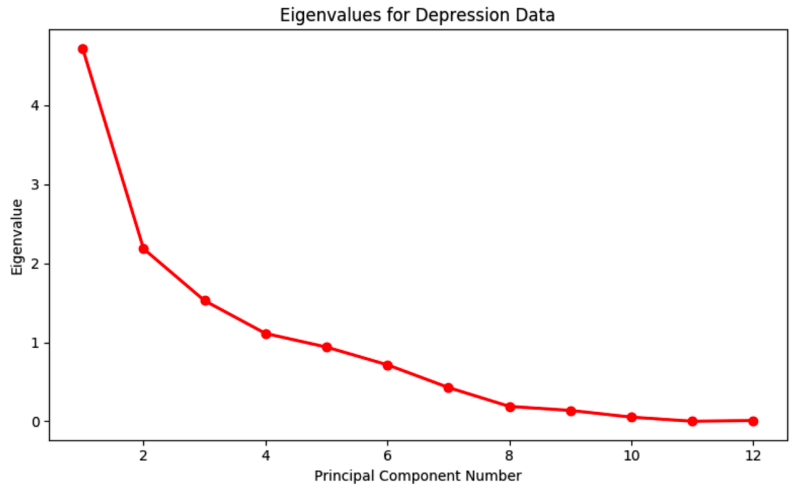
Please note that assignments in this class may be submitted to www.turnitin.com, a web- based anti-plagiarism system, for an evaluation of their originality.

Question: Apply PCA analysis on x_1, ..., x_n of Classification Data2.xlsx and then select a few principal components based on the scree plot. Also, clarify which variables have a high correlation (>0.5) with the first and second principal components.

Answer:

By using PCA analysis, I got following plot.





We can see 6 features can capture over 95% of the variance within the dataset, so we choose 6 principal components.

And if we use R rule, choose eigenvalues > 1. We can choose 4 principal components.

And then calculate the eigenvalues for the principal components.

```
[4.72240243e+00 2.18412049e+00 1.52731938e+00 1.11223615e+00
 9.39673439e-01 7.15172523e-01 4.28068625e-01 1.88564075e-01
 1.37875499e-01 5.30488130e-02 9.70990108e-04 9.93368639e-03]
```

We can get following result:
PC1 = 4.722
PC2 = 2.184
And following are the eigenvectors for the first 2 principal components:

```
[[-0.41681538 -0.13984697 -0.23640314 -0.13205569  0.02891208  0.10470243
   0.15729023 -0.4514806   0.63446387 -0.22180217 -0.19806259 -0.04516372]
 [-0.30228256  0.28937891  0.14707551  0.42647581  0.16719785  0.17331862
  -0.30214789 -0.00435836  0.00252199  0.02225754 -0.00231936 -0.68670105]]
```

Since we want variables have correlation > abs(0.5) (consider positive and negative correlation here):
For 1st principal components:
Abs(0.5) / sqrt (4.722) = ± 0.23. Therefore, X1, X3, X8, X9's correlation > abs(0.5)

And for 2nd principal components:
Abs(0.5) / sqrt (2.184) = ± 0.338. Therefore, X4, X12's correlation > abs(0.5)

If we only consider the variables have correlation > 0.5 (only positive correlation):
For 1st principal components:
0.5 / sqrt (4.722) = 0.23. Therefore, X9's correlation > 0.5

And for 2nd principal components:
0.5 / sqrt (2.184) = 0.338. Therefore, X4's correlation > 0.5