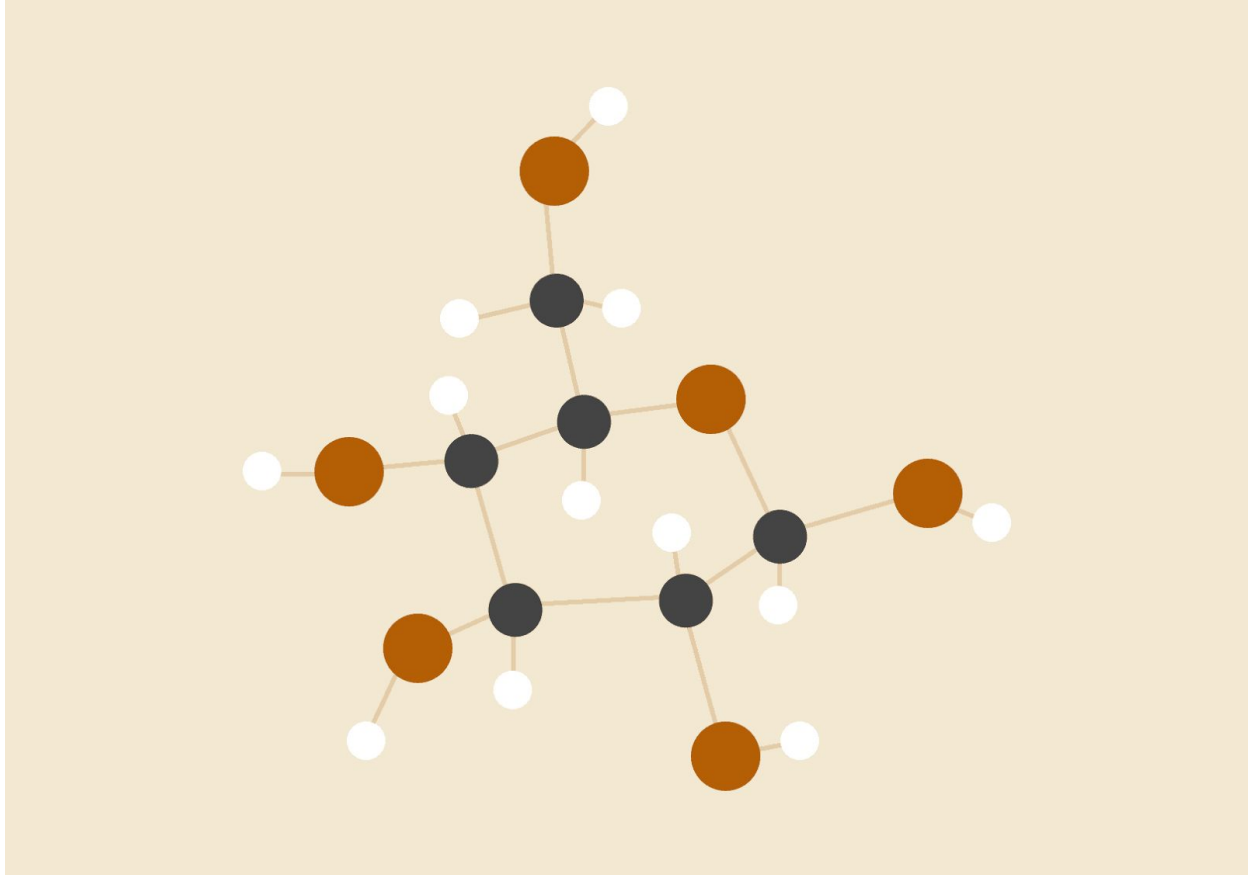


Project -1 REPORT

Introduction to Machine Learning - CSE574



Jayant Solanki

UBIT Name: jayantso

Person Number: 50246821

Fall 2017 CSE

INTRODUCTION

This report contains the implementation report for the Task 1 , 2 and 3 of the Project 1. The report talks about the implementation in chronological fashion, right from the data acquisition from the DataSet to the computation of loglikelihood for the given 4 feature columns mentioned in the project description pdf.

OBJECTIVES ACHIEVED

I have used different statistical tools based upon numpy and python. I calculated, mean, variance, standard deviation on 4 separate feature columns which are **CS Score (USNews)**, **Research Overhead %**, **Admin Base Pay\$** and **Tuition(out-state)\$** of the university data set. As per project requirement I also calculated covariance and correlation matrices for those 4 feature columns. Scatter plots graphs have been plotted for each feature pairs. In the end I calculated univariate loglikelihood and multivariate loglikelihood.

IMPLEMENTATION

Using **openpyxl** python library all the 4 relevant columns from the “**university data.xlsx**” were read and stored into a ‘**data**’ numpy array of 49x4 shape.

Task -1 : Calculated mean, variance, and standard deviation on the 4 respective columns in the **data** array using **numpy.mean()**, **numpy.var()** and **numpy.std()**. Stored the results in the respective mu1, mu2....m4, var1,...var4 and sigma1...sigma4 variables.

Task -2 : Calculated covariance matrix and correlation matrix on the 4 columns using **numpy.cov()** and **numpy.corrcoef()** functions. Here the data array was transposed before using it as an input for the two functions. Results were stored inside the covarianceMat and correlationMat. Based upon the correlation matrix I found that columns **CS Score (USNews)** and **Research Overhead %** were most correlated variable pairs and columns **CS Score (USNews)** and **Admin Base Pay\$** were least correlated variable pairs. Below is the scatter plot graphs between each pair of feature columns in the respective Figure 1, 2 and 3.

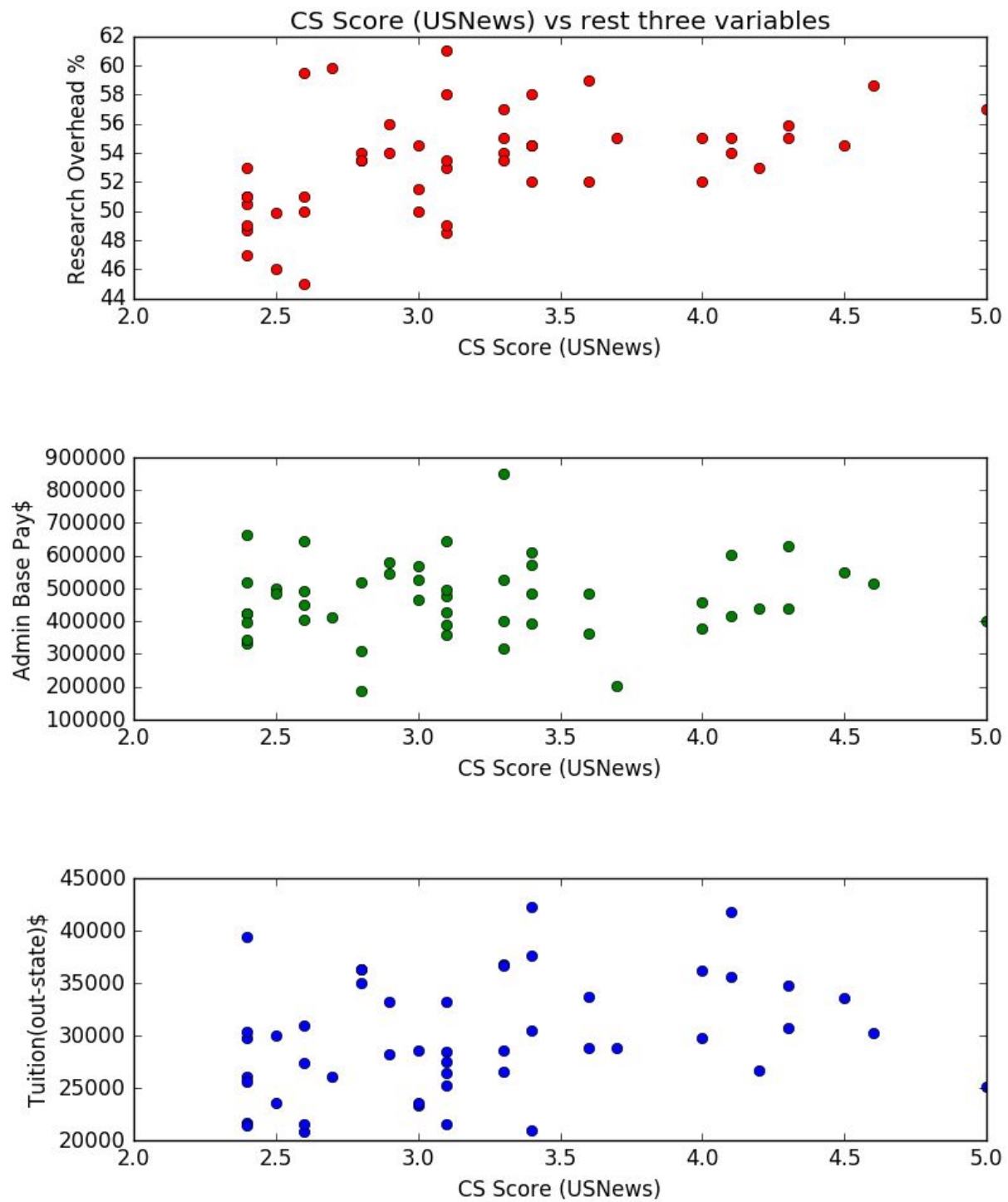


Figure 1

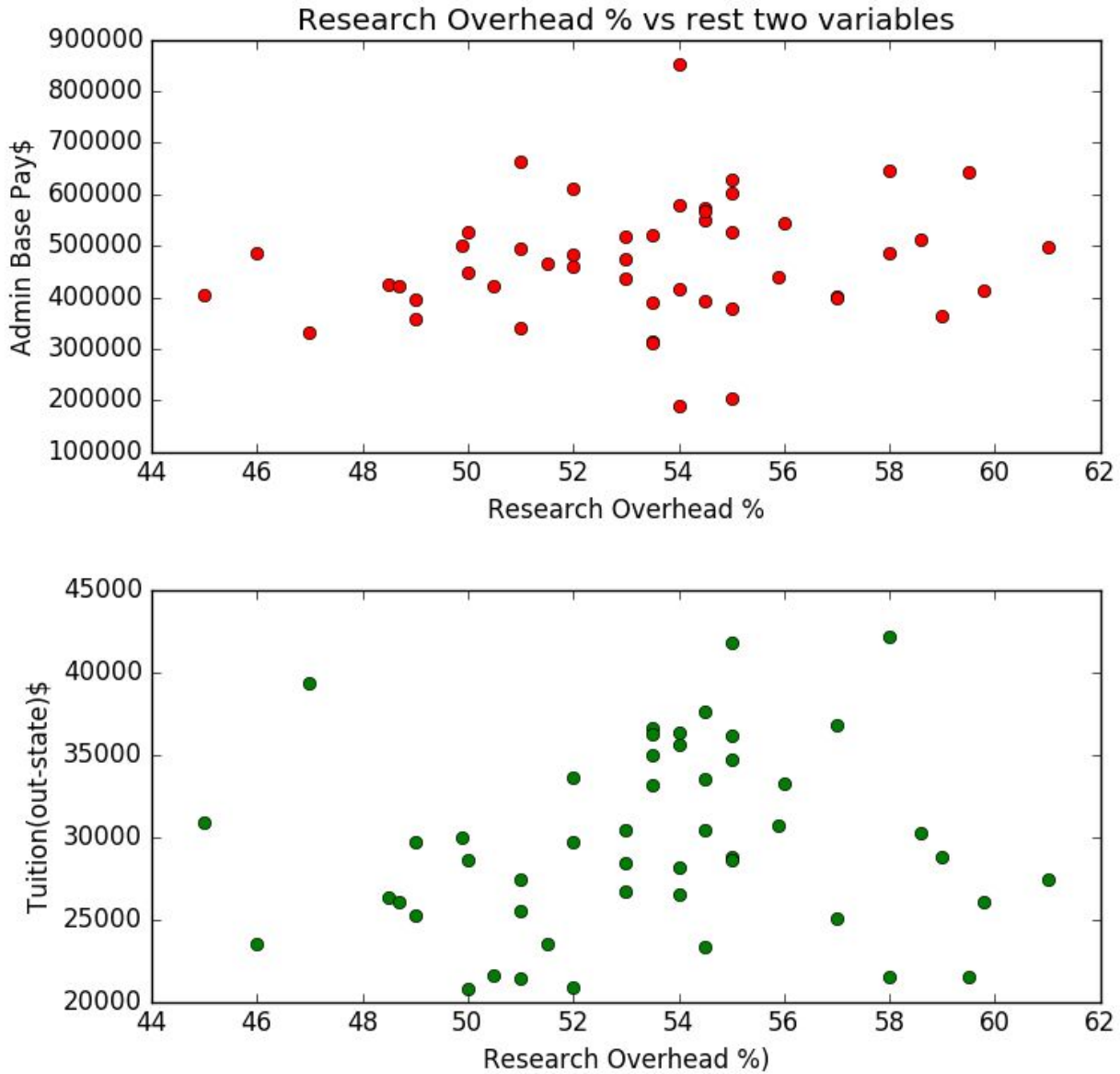


Figure 2

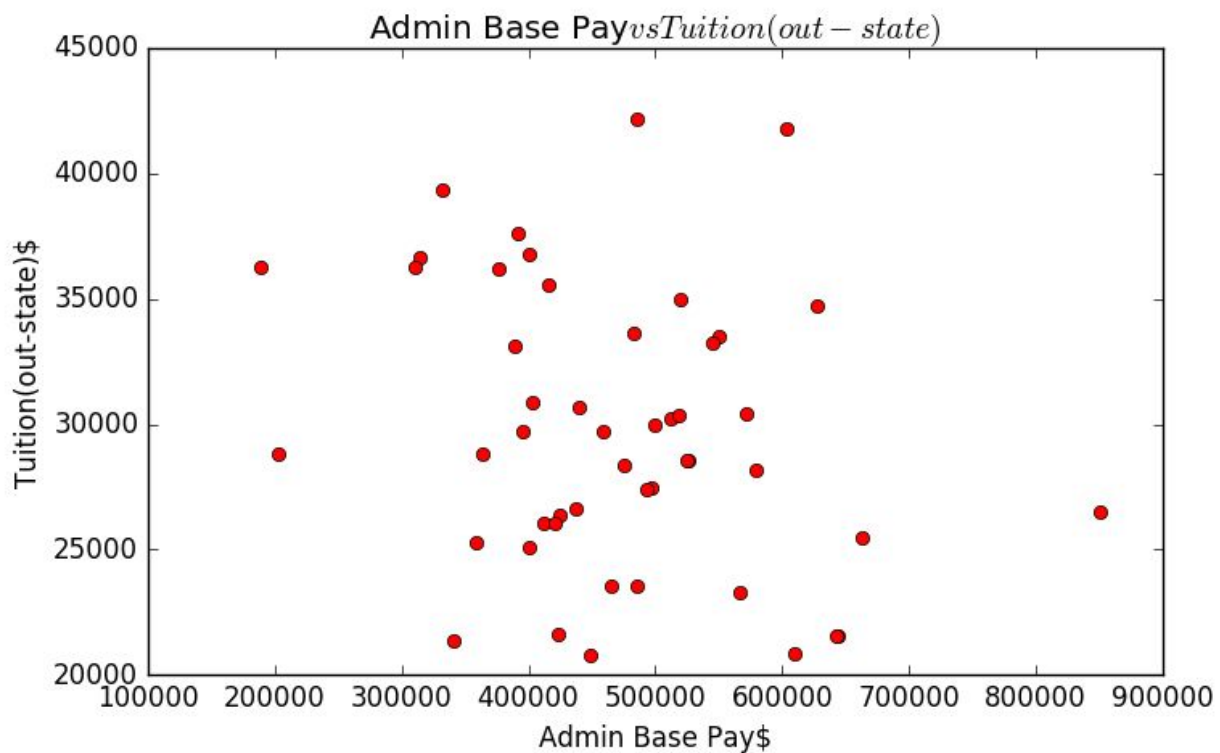


Figure 3

Task -3 : Assuming that each column is normally distributed and that they are independent of each other I calculated the log-likelihood of each 4 columns using `scipy.stats.norm()` function, took the sum of their logs for each column data point and added all four values to find the overall log-likelihood.

Now considering the 4 feature columns are not independent, I calculated the multivariate log-likelihood. I used the probability distribution function for multivariate features. Created a mean vector, and by using the covariance matrix I calculated the multivariate log-likelihood for 4 feature variables.

MATERIALS

1. ProbabilityConcepts.pdf
2. Project1Desc.pdf
3. DataSet.zip

RESULTS

```
Ubit Name = jayantso
personNumber = 50246821
mu1 = 3.214
mu2 = 53.386
mu3 = 469178.816
mu4 = 29711.959
var1 = 0.448
var2 = 12.588
var3 = 13900134681.7
var4 = 30727538.733
sigma1 = 0.669
sigma2 = 3.548
sigma3 = 117898.832
sigma4 = 5543.243
covarianceMat =
[[ 4.575e-01  1.106e+00  3.880e+03  1.058e+03]
 [ 1.106e+00  1.285e+01  7.028e+04  2.806e+03]
 [ 3.880e+03  7.028e+04  1.419e+10 -1.637e+08]
 [ 1.058e+03  2.806e+03 -1.637e+08  3.137e+07]]
correlationMat =
[[ 1.   0.456  0.048  0.279]
 [ 0.456  1.   0.165  0.14 ]
 [ 0.048  0.165  1.   -0.245]
 [ 0.279  0.14 -0.245  1.   ]]
logLikelihood = -1315.099
multilogLikelihood = -1304.778
```

SOFTWARE/HARDWARE USED

- Sublime Text 3, Python 3 Environment based upon Anaconda, Ubuntu 16 System, Intel core i3 processor.
- Python libraries: numpy, openpyxl, matplotlib and scipy

REFERENCES

1. Ublearns
2. Stackoverflow.com
3. Python, Numpy and Matplotlib documentations.