University at Buffalo
The State University of New York

INTRODUCTION TO MACHINE LEARNING
CSE 574

FALL-2016

PROJECT-1 REPORT

SUBMITTED BY:
ARUN CHANDRA PENDYALA(50207136)

REPORT

1.INTRODUCTION:

In this project, various statistical measures such as mean, variance, standard deviation, covariance and correlation coefficients have been evaluated using python code for the given sets of data and the Bayesian network, which shows the probabilistic relationships between various data sets, has been constructed such that the log likelihood is improved from the case in which each variable is independent of each other.

The four variables which form the given set are CS score, Research overhead, Admin base pay, Tuition. For the purpose of finding the Loglikelihood, each variable is assumed to follow normal distribution.

Bayesian network for a given set of variables attempts to maximize the Loglikelihood for the observed data. Linear Gaussian model is assumed for the construction of Bayesian network.

2.IMPLEMENTATION:

2.1.OVERVIEW:

Data corresponding to each variable is extracted from the university data excel sheet and the mean, variance, standard deviation, covariance matrix and correlation matrix are computed using Numpy module. Scatterplots are plotted for various combinations of variables. The data matrix is formed by merging the variable lists as the column vectors. The Loglikelihood is found out by assuming separate cases: i)the variables are independent and ii) the variables form a Bayesian network.

2.2.BAYESIAN NETWORK:

There are 4 variables : X1 , X2 , X3 , X4 corresponding to CS score , Research overhead, Administrator base pay, Tuition. The Bayesian network can be formed by trying out various combinations of these variables such that a directed acyclic graph is formed from these variables/nodes. The Bayesian network formed is expected to improve the Loglikelihood than in the case in which the variables are assumed to be independent of each other.

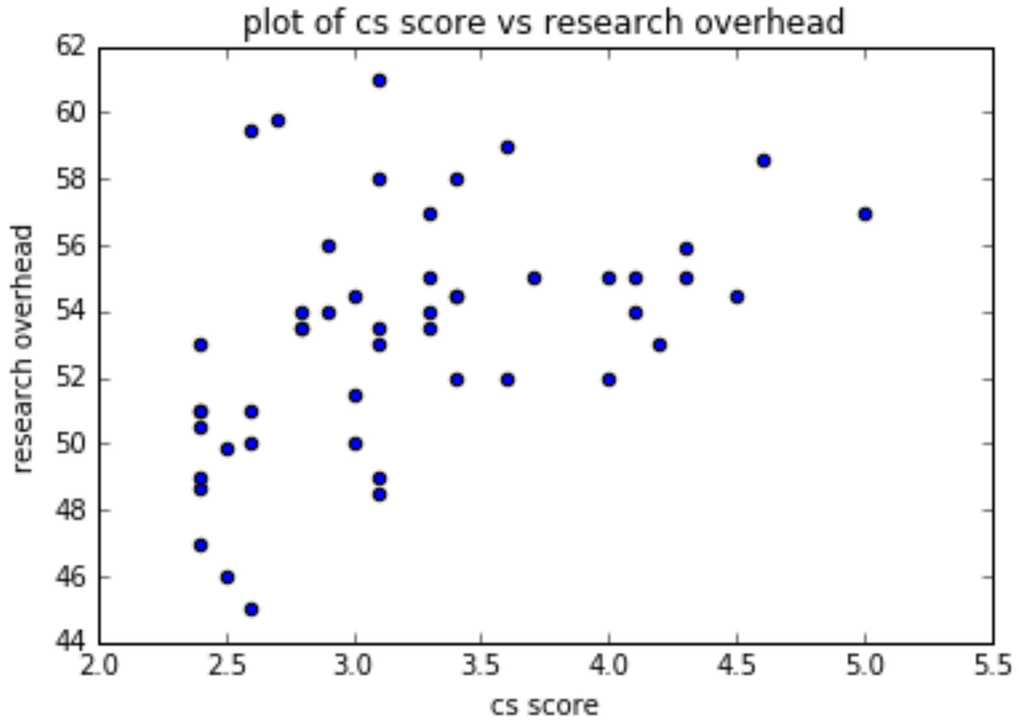
The variables are as follows:
X1 – CS score
X2 – Research overhead
X3 – Administrator base pay
X4 - Tuition

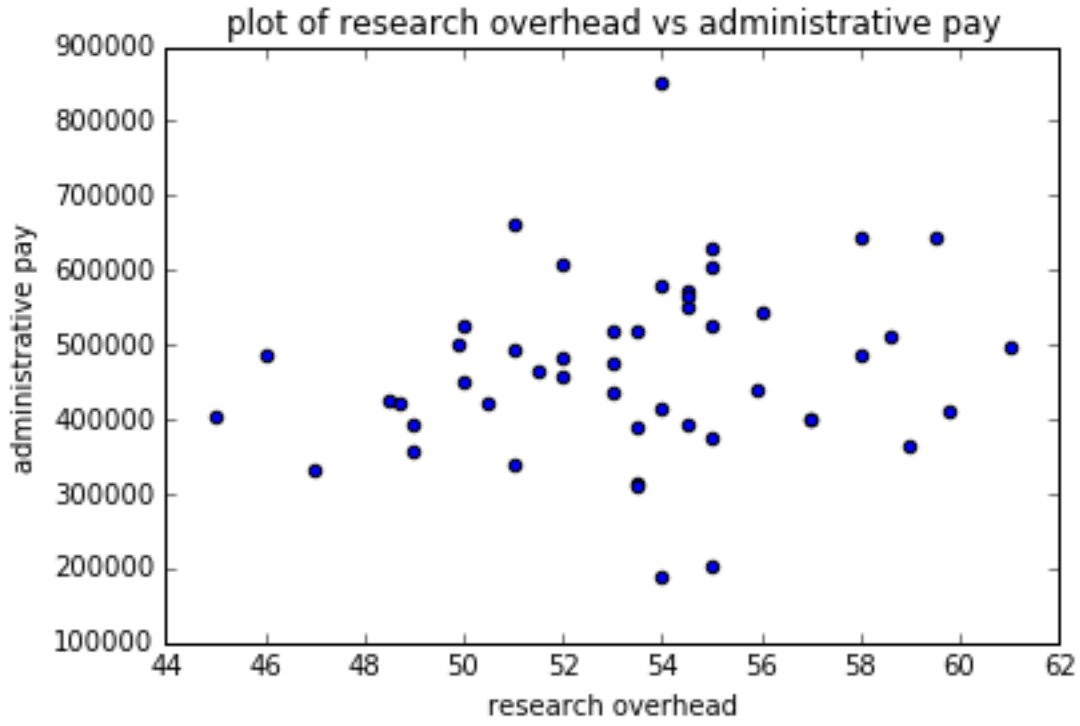
The results for Bayesian network which has improved loglikelihood have been discussed in the next section.

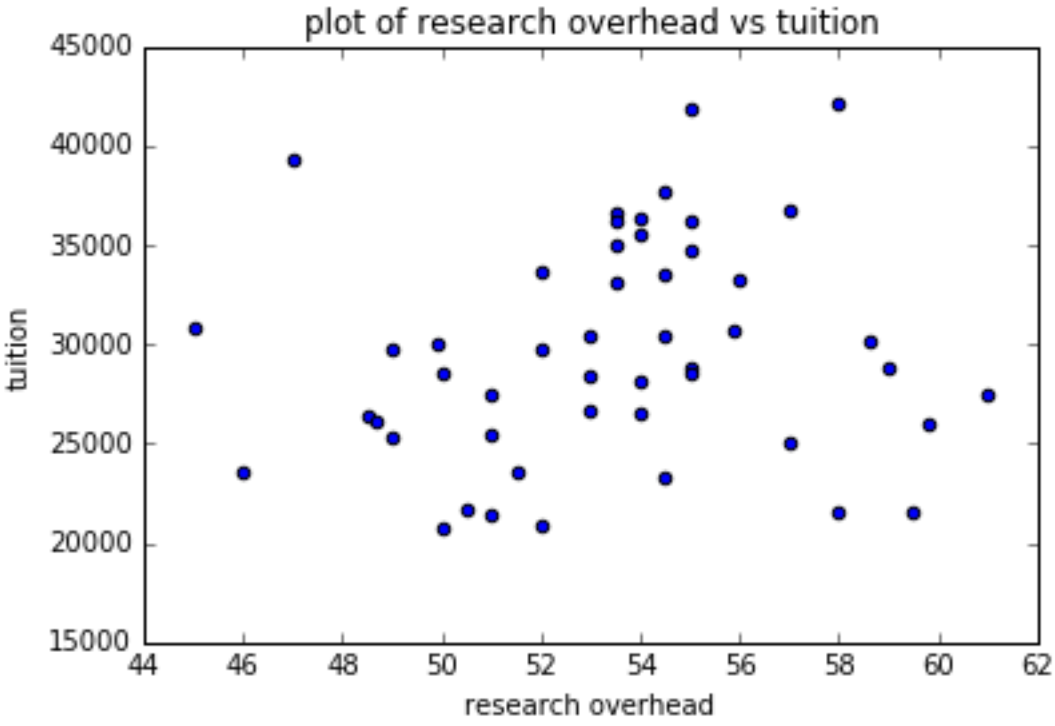
3.RESULTS:

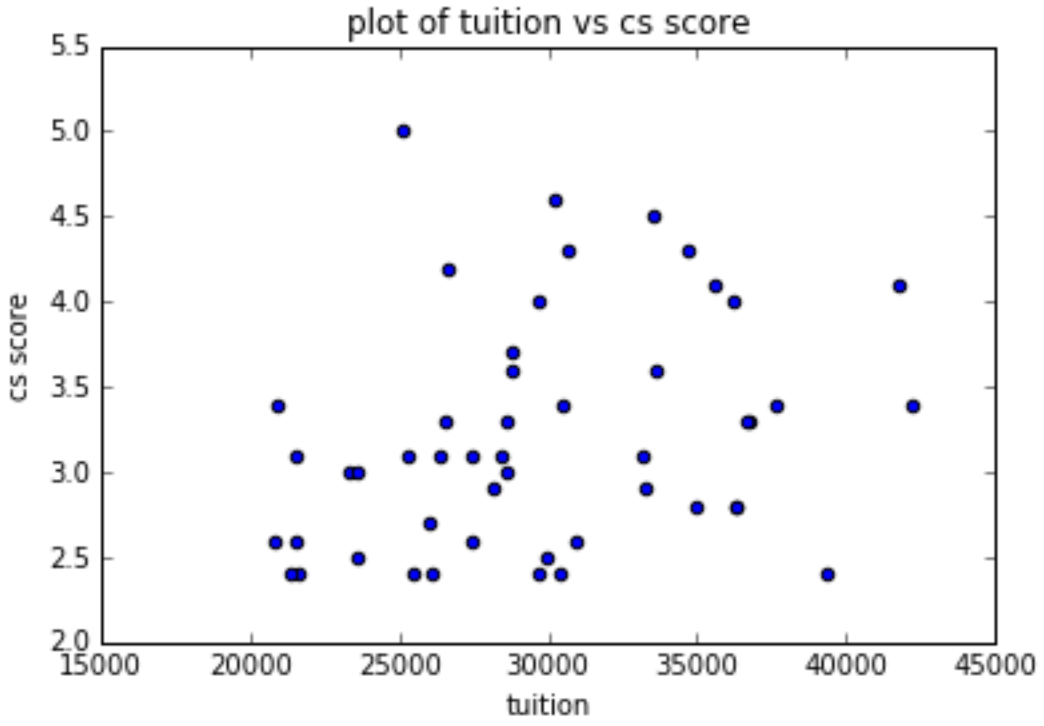
3.1:OUTPUT OF PYTHON CODE:

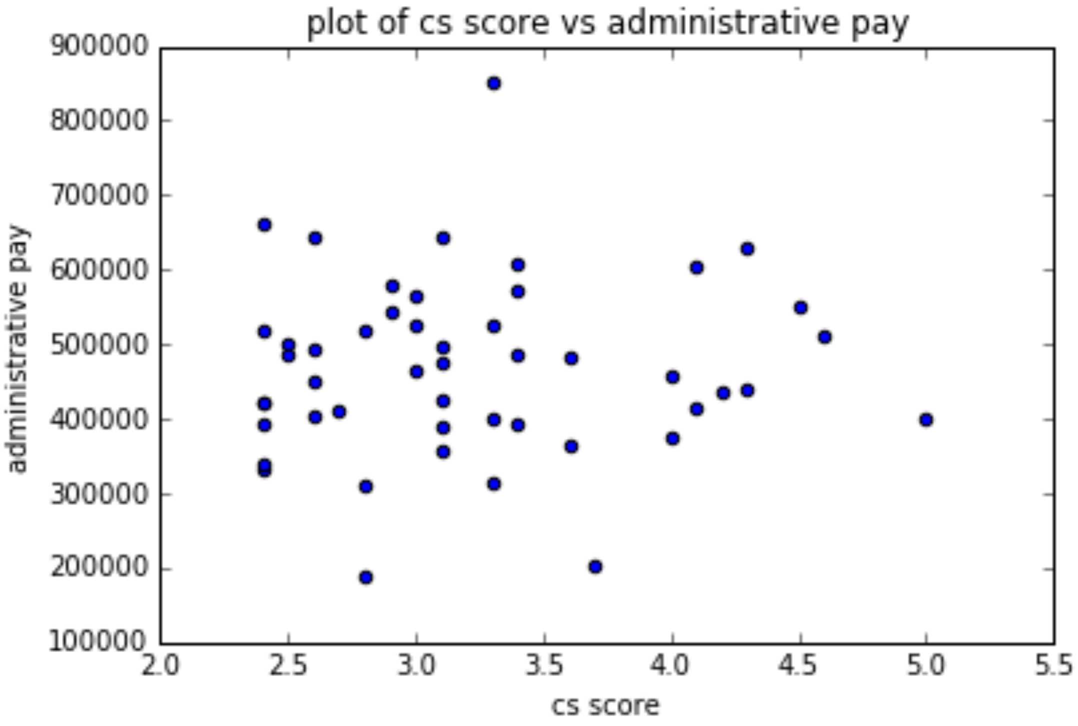
```
UBitName = apendyal
personNumber = 50207136
mu1 = 3.21
mu2 = 53.39
mu3 = 469178.82
mu4 = 29711.96
var1 = 0.45
var2 = 12.59
var3 = 13900134681.70
var4 = 30727538.73
sigma1 = 0.67
sigma2 = 3.55
sigma3 = 117898.83
sigma4 = 5543.24
covarianceMat =
[[  4.60000000e-01   1.11000000e+00   3.87978000e+03   1.05848000e+03]
 [  1.11000000e+00   1.28500000e+01   7.02793800e+04   2.80579000e+03]
 [  3.87978000e+03   7.02793800e+04   1.41897208e+10  -1.63685641e+08]
 [  1.05848000e+03   2.80579000e+03  -1.63685641e+08   3.13676958e+07]]
correlationMat =
[[ 1.     0.46   0.05   0.28]
 [ 0.46   1.     0.16   0.14]
 [ 0.05   0.16   1.    -0.25]
 [ 0.28   0.14  -0.25   1.  ]]
logLikelihood = -1315.10
BNgraph =
[[0 0 0 0]
 [1 0 0 0]
 [1 0 0 0]
 [1 0 0 0]]
BNlogLikelihood = -1307.84
```

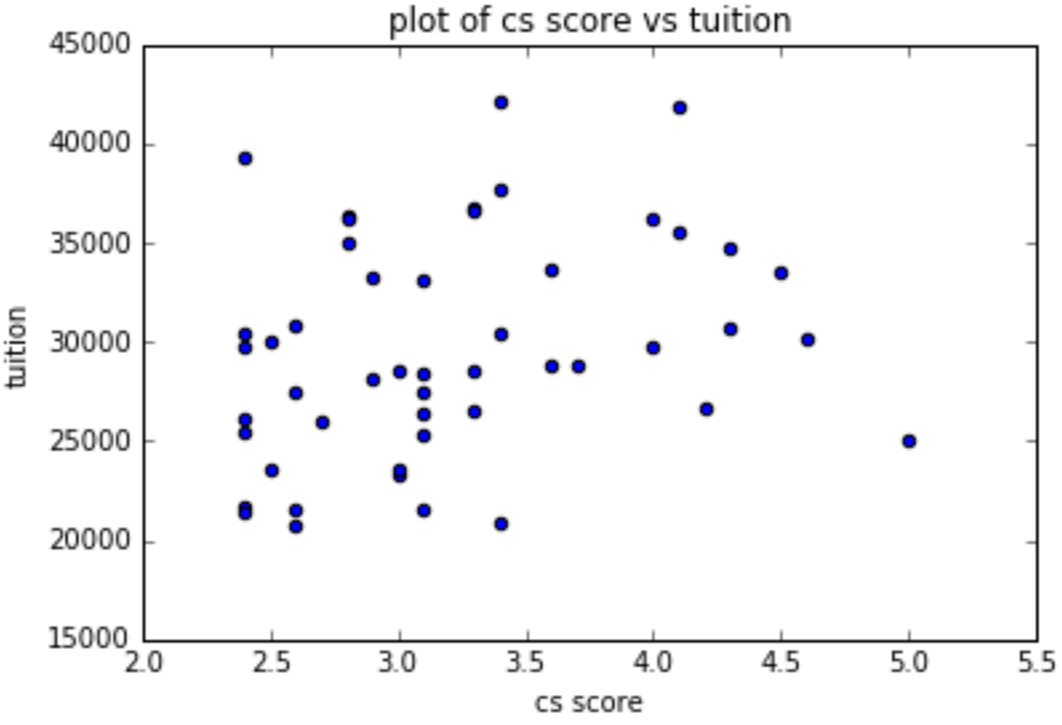

plot of cs score vs research overhead
research overhead
62
60
58
56
54
52
50
48
46
44
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
cs score


plot of research overhead vs administrative pay
administrative pay
900000
800000
700000
600000
500000
400000
300000
200000
100000
44
46
48
50
52
54
56
58
60
62
research overhead


plot of research overhead vs tuition
tuition
research overhead
45000
40000
35000
30000
25000
20000
15000
44
46
48
50
52
54
56
58
60
62

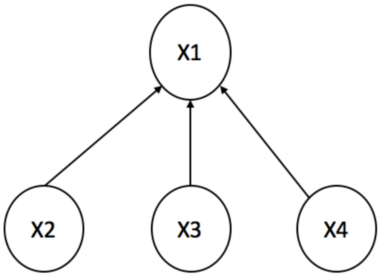

plot of tuition vs cs score
cs score
tuition
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
15000
20000
25000
30000
35000
40000
45000


plot of cs score vs administrative pay
administrative pay
cs score
900000
800000
700000
600000
500000
400000
300000
200000
100000
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5


plot of cs score vs tuition
tuition
cs score
45000
40000
35000
30000
25000
20000
15000
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

Highest correlation is for the variable pair – X1 AND X2 = 0.46
Lowest correlation is for the variable pair – X4 AND X3 = -0.25

The constructed Bayesian network which corresponds to the results – BNgraph and BNloglikelihood = -1307.83 is as follows:

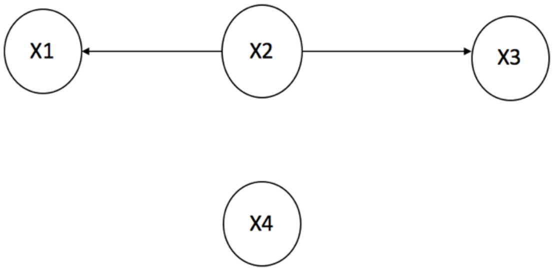


X1 – CS score
X2 – Research overhead
X3 – Administrator base pay
X4 – Tuition

The **observation** that can be made from the graph is that the Research overhead , Administrative basepay and tuition all seem to influence CS score for a university which seems quite plausible.

The following Bayesian networks have also been tried out:

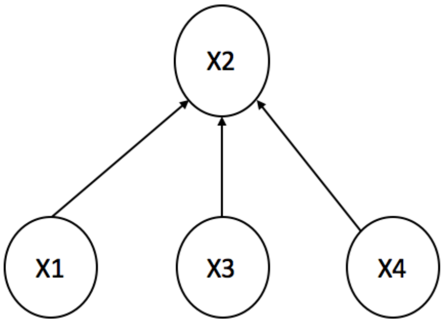
1)The following network has been inferred from the plots of pairwise data[1]. The research overhead seems to influence CS score and administrator pay. From the scatter plots, it appears that the tuition is independent of the other variables.



BNloglikelihood = -1308.72

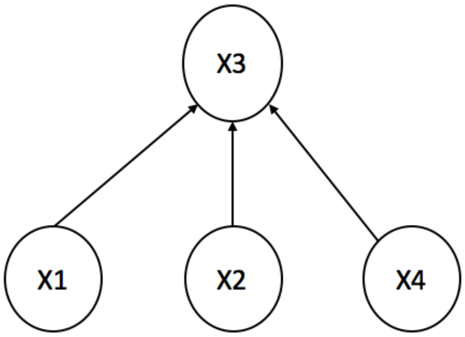
2)The following graphs are the altered versions of the constructed Bayesian network:

2.1)



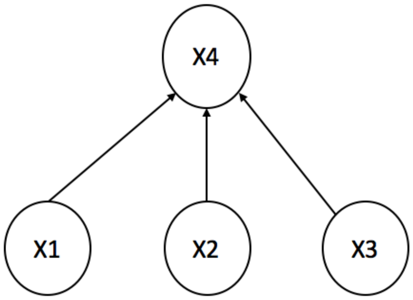
BNloglikelihood = -1308.66

2.2)



BNloglikelihood = -1312.46

2.3)



BNloglikelihood = -1311.16

4.REFERENCES:

[1] Interpretation of scatterplot: http://www.uow.edu.au/student/qualities/statlit/module3/5.4interpret/index.html