PROJECT REPORT

CSE 474/574: Introduction to Machine Learning

Instructor: Sargur N. Srihari

Arun Sharma

UB Person: 50206920

Overview

We have given a dataset of universities having 9 columns each containing information about university and there rating according to it. Only 4 columns were used for this project which are CS score, Research Overhead, Admin Base and Tuition. Simple operation like mean standard deviation, variance were calculated and computed for finding correlation and covariance of the data. We find some interesting results when observing the correlation matrix. Also further for Task 3 we computed log-likelihood of the data but for better score we attempt to make a Bayesian network and try different combinations.

Bayesian Networks and their Likelihood Value

Since log likelihood was computed for every column independently, it is out intuition to make a dependent log likelihood model which might give us relatively better results than existing independent model.

Let us see how Probabilistic model like Bayesian network gives us better results by examine some combinations of nodes.

Firstly, we examined some interesting combination through correlation graph:

```
[[ 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]]
```

As we can see positive correlation values (except the diagonal values) are 0.456, 0.048, 0.279 and 0.165. Hence they are more dependent to each other, therefore these nodes should be involved in the Bayesian network.

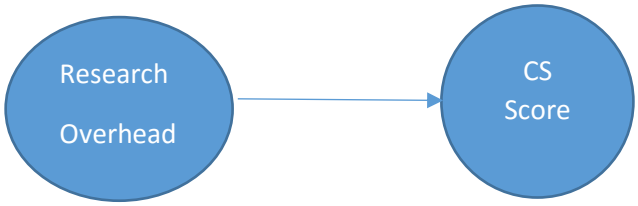


Fig: When Research Overhead is only parent of CS Score

The value 0.456 involves CS score and Research Overhead depicts they are strongly correlated to each other. For example if we take log likelihood P(CS Score|Research Overhead) we get -44.153

log likelihood, hence from that we can conclude that those nodes having strong correlation have better log likelihood.

Similarly, we can compute for every parent to cs score and we find log likelihood such as P(CS Score | Admin base pays) giving value of log likelihood of -49.13 etc.

Finally, we get a best model for giving the best log likelihood -46.602. Detail results are given in Results and graph section

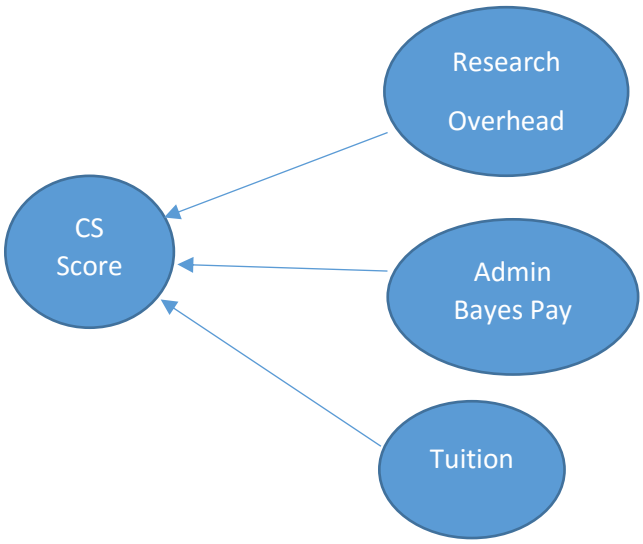


Fig: Bayesian Network used in this project

Methodology

The project methodology is as follows:

1) Computing mean, variance and standard deviation
2) Computing covariance and correlation matrix
3) Computing pdf of each and every column INDEPENDENTLY
4) Taking log of the likelihood of every column of pdf and sum it
5) After we compute the log likelihood with the help of Bayesian model by trying different combinations of parent and child node
6) Finally computed the model giving us the best score.
7) Comparison and analysis of Bayesian network with independent model.

Implementation

For implementation we used Anaconda IDE for computation along with packages like numpy, scipy and xlrd. We used windows 10 as operating system. We also used matplot for computing the scatterplot graph for representing each and every column (Since matplot library was not able to execute in timberlake server, I include png files of the plot for the scatter plot as instructed by TA's).

Various Bayesian Network was computed, but the highest score (best score) was -42.602 when CS Score was a child node and every other are the parent node.

Results and Graphs:

Results regarding the project is described below:

1) Individual mean variance and standard deviation

	CS Score	Research Overhead	Admin Base Pay	Tuition
Mean	3.214	53.386	469178.816	29711.959
Variance	0.448	12.588	13900134681.7	30727538.7
Standard Deviation	0.669	3.548	117898.832	5543.243

2) Covariance matrix

[[4.58000000e-01 1.10600000e+00 3.87978200e+03 1.05848000e+03]

 [1.10600000e+00 1.28500000e+01 7.02793760e+04 2.80578900e+03]

 [3.87978200e+03 7.02793760e+04 1.41897208e+10 -1.63685641e+08]

 [1.05848000e+03 2.80578900e+03 -1.63685641e+08 3.13676958e+07]]

3) CorrelationMat

[[1. 0.456 0.048 0.279]

 [0.456 1 0.165 0.14]

 [0.048 0.165 1 -0.245]

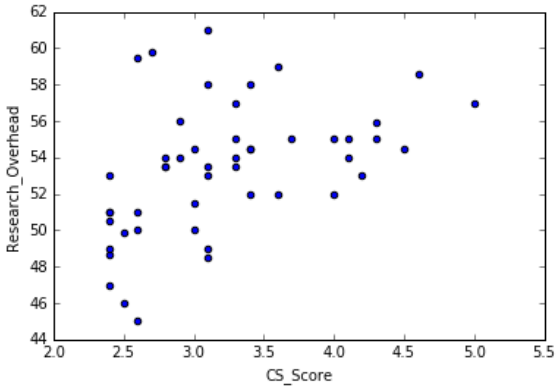
 [0.279 0.14 -0.245 1.0]]

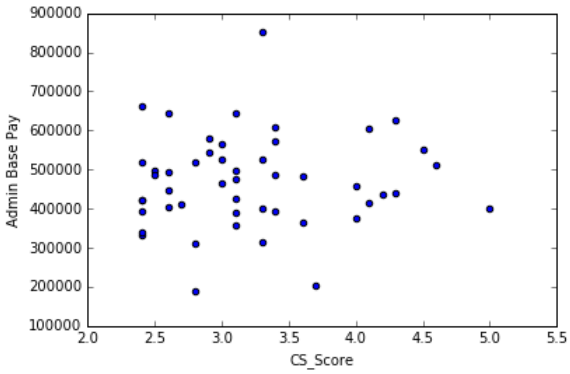
4) Likelihood = -1315.009

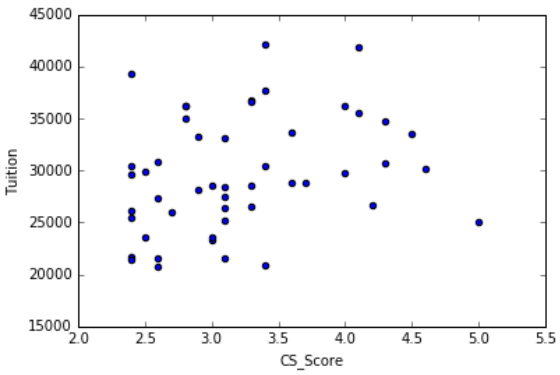
5) BNGraph = [[0 0 0 0] [1 0 0 0] [1 0 0 0] [1 0 0 0]]

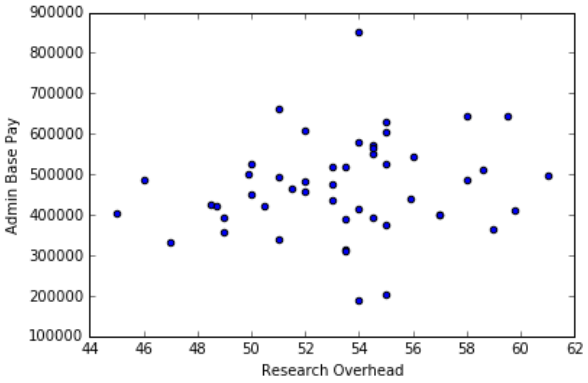
6) BNLoglikelihood = -42.602

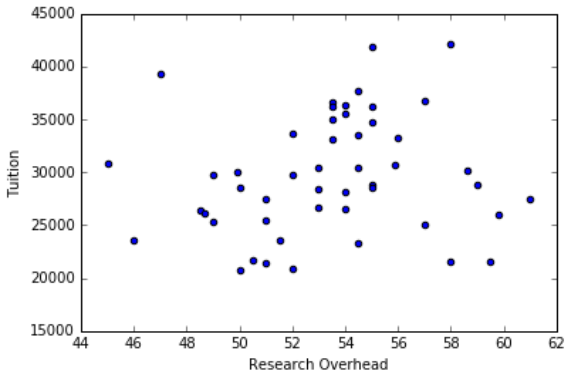
While calculating the graphs we find that the dots which are very close to each other are the values which highly contribute to the correlation. For example, in the first graph CS_Score and Research Overhead are most correlated so values in dots are very close to each other. Likewise, the other graphs contribute to

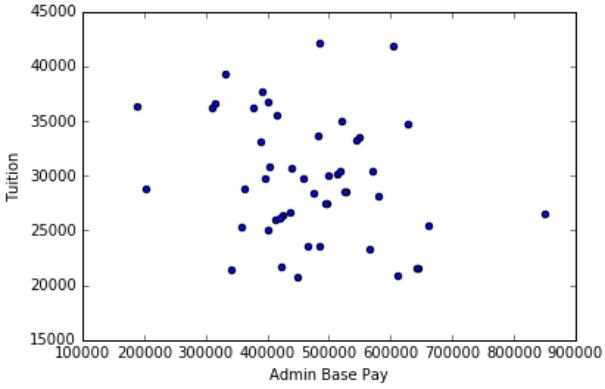












Conclusion

In this project we find some interesting relationship of the nodes in the Bayesian network. Also we related through correlation between the nodes with their log likelihood which was giving higher value than it was calculated independently. However, there were many combinations which weren't used since we are calculating and designing the Bayesian network through their correlation value. Also many interesting conditional probability is yet to be calculated which can give further more information.