Project Report 1

Probability Distributions and Bayesian Networks

CS 574 – Fall 2016

Raj Jaysukh Patel:

(UBIT Name: rajjaysu UBIT : 50208278)

1. Introduction

Probability Distribution is a function of a discrete variable whose integral over any interval is the probability that the random variable specified by it will lie within that interval. Bayesian Network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). The project was about analyzing data given and understanding the dependencies among the variables to design a Bayesian network to get better log likelihood then independent variables.

2. Data

The data given in this project is about universities, which has four variables which are ranking in CS, research overhead, admin pay and tuition fees. The basic task was to find the mean, variance and standard deviation of the given variables for further calculations.

3. Analysis and Results

Co-relations and Covariance

The co-relation matrix and covariance matrix was obtained as a part of task 2, to understand the variance and relation between all the variables. From the results it is quite clear that the CS Score and Research Overhead are most co-related variables with coefficient 0.456 and CS Score and Admin Base Pay has least co-relation with coefficient of 0.048. The co-relation matrix and pair wise graph for above variables are as shown below.

Table 1 Co-relation Matrix

	CS Score	Research Overhead	Admin Base Pay	Tuition
CS Score	1.0	0.456	0.048	0.279
Research Overhead	0.456	1.0	0.165	0.14
Admin Base Pay	0.048	0.165	1.0	-0.245
Tuition	0.279	0.14	0.245	1.0

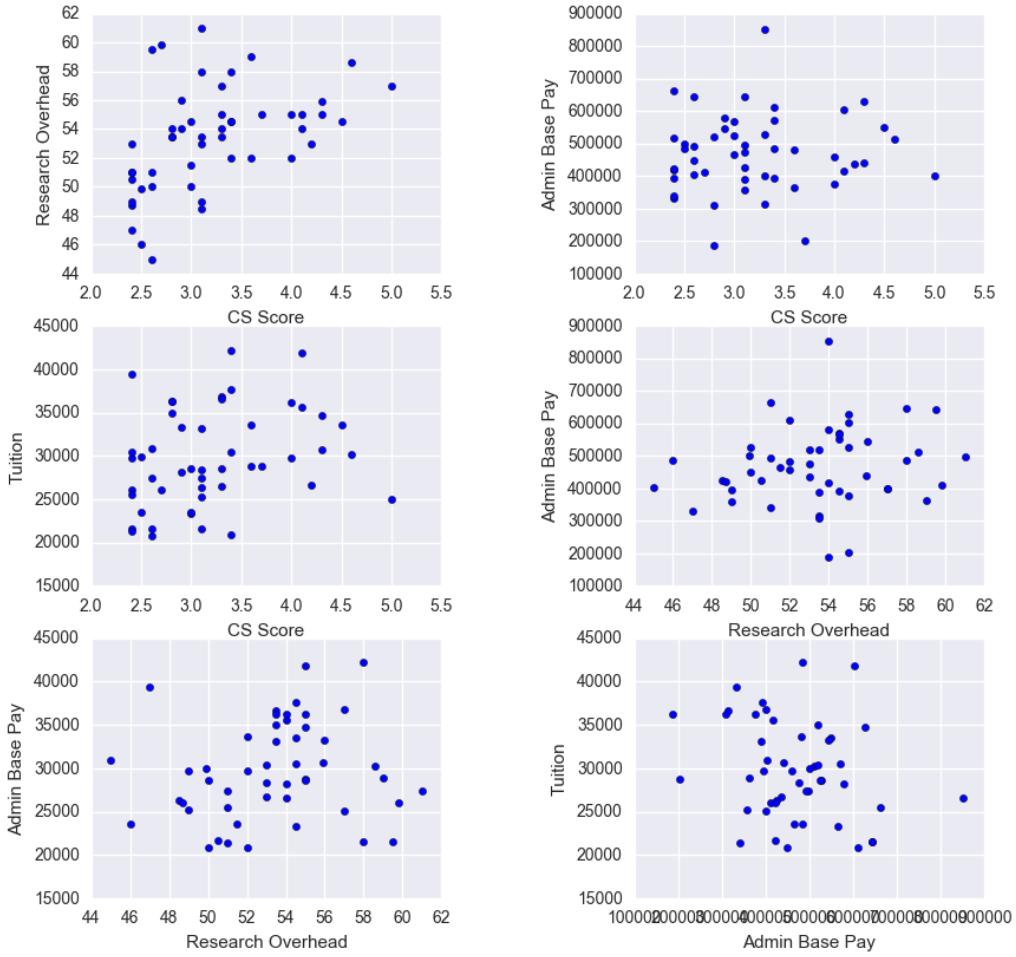

Figure 1 Pair Wise Plotting

Bayesian Network

Bayesian Network are basically Directed Acyclic Graph with edges showing the dependencies of different variables. This gives the parent child relation, for example the edge from A to B denotes that A is parent and B is its child so, B is dependent on A. Looking at the co-relation matrix and pair wise graph of all variables it can be assumed that CS Score is dependent on two parent nodes: Research Overhead, Tuition. Research Overhead is dependent on one parent nodes: Admin Base Pay Tuition is dependent on 2 parent nodes: Research Overhead, Admin Base Pay. This makes sense as Research Overhead and Tuition will increase the CS Score. Admin Base Pay will also increase the Research Overhead. Research Overhead and Admin Base Pay will increase the Tuition. Based on this we derive Bayesian Network Graph as follows.

Table 2 Bayesian Network Graph Matrix

	CS Score	Research Overhead	Admin Base Pay	Tuition
CS Score	0	0	0	0
Research Overhead	1	0	0	1
Admin Base Pay	0	1	0	1
Tuition	1	0	0	0

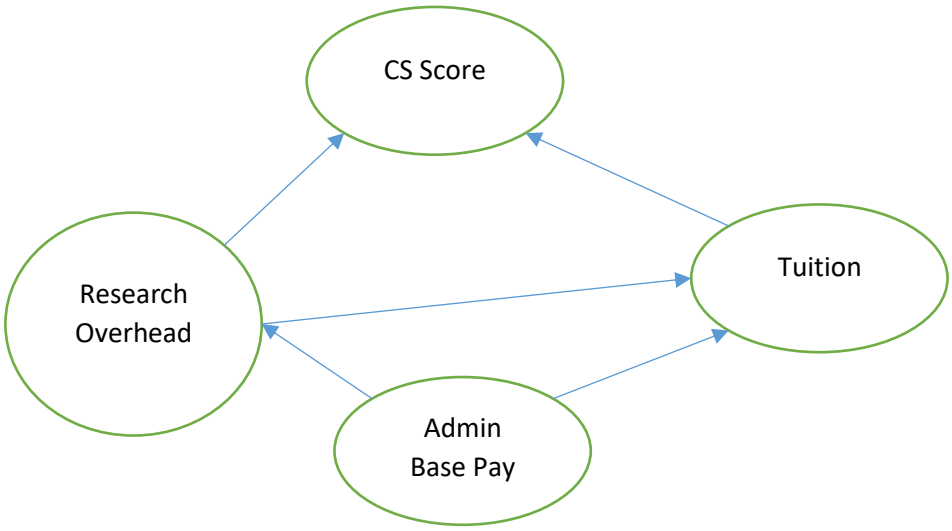

Figure 2 Bayesian Network

From the above graph we can derive multi variant probability distribution as follows:

Consider,
CS Score as C
Research Overhead as R
Admin Base Pay as A
Tuition as T

P(C,R,A,T) = P(C/R,T) * P(R/A) * P(A) * P(T/R,A)
BNLogLikelihood = (log(P(C/R,T)) + (log(P(R/A)) + (log(P(R)) + (log(P(T/R,A))

Log (P (Y/ $X_1, \dots X_k$))
= $\sum_{n=1}^{N}[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\beta_0 x_0[n] + \beta_1 x_1[n] + \cdots + \beta_k x_k[n]\text{-}y[n])^2]$

Where k = number of parents,
Y is child
X is parent
$x_0[n]$ = 0

BNLogLikelihood for above Bayesian Network is -1305.884 (obtained by the python program) which is greater than normal logLikelihood which is -1315.099.

4. Interesting Conditional Probability

LogLikelihood(P(C/R,T) = log(P(C/R,T)) = -43.605424818

LogLikelihood(P(R/A)) = log(P(R/A)) = -130.916887418

LogLikelihood(P(R)) = log(P(R)) = -641.729515103

LogLikelihood(P(R)) = log(P(T/R,A)) = -489.632519451

5. Conclusion

As per the results obtained for different multi variant probability distribution and conditional probability we can conclude that Bayesian Network can be developed for given data on basis of dependencies and logLikelihood can obtained for the Bayesian Network will optimal than that obtained normally.

6. References

- Probability Theory - Pattern Recognition and Machine Learning by Christopher M. Bishop
- Bayesian Network - https://en.wikipedia.org/wiki/Bayesian_network
- Supplementary Material for Bayesian Network Problem