Constructing an optimal Bayesian Network Model for a given Data Set

Vaibhav Lella, Ubit Name : vaibhavl , Person Number:50169859

ABSTRACT

Bayesian Networks have played a significant role in the field of AI over the last decade. It provides a way to represent knowledge in an uncertain domain and a way to reason about this knowledge. The task at hand is to construct a Bayesian network for the given multivariate dataset which results in a very high log likelihood and helps capture the relation between variables present in the dataset.

1. INTRODUCTION

In this project we design a Bayesian network model for the multivariate dataset given. First we determined the LogLikelihood of the dataset using already available mathematical formulas Later using causal reasoning and by analyzing the correlation and covariance matrix values we constructed directed acyclic graphs that sort of fit the data and capture the relationship among the different variables. For the different possible Bayesian graphs we calculated the log likelihood values until we came up with an optimal value greater than the one calculated earlier. We used MatlabR2012a for all the mathematical calculations.

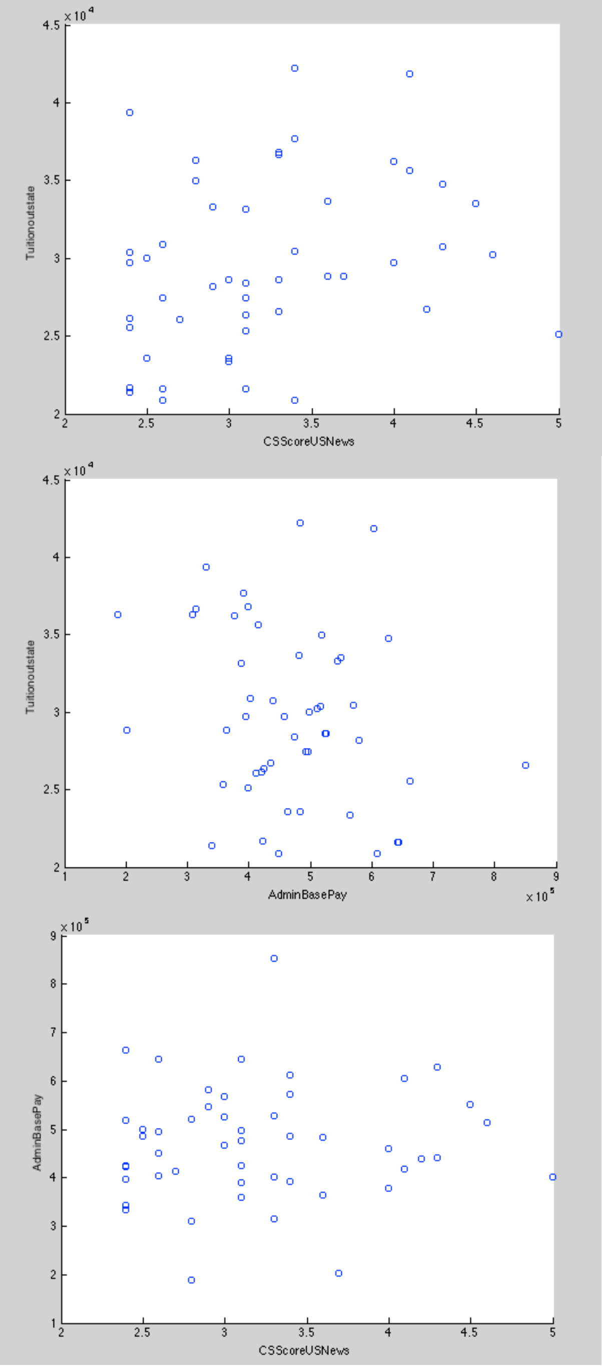
2. THE DATASET

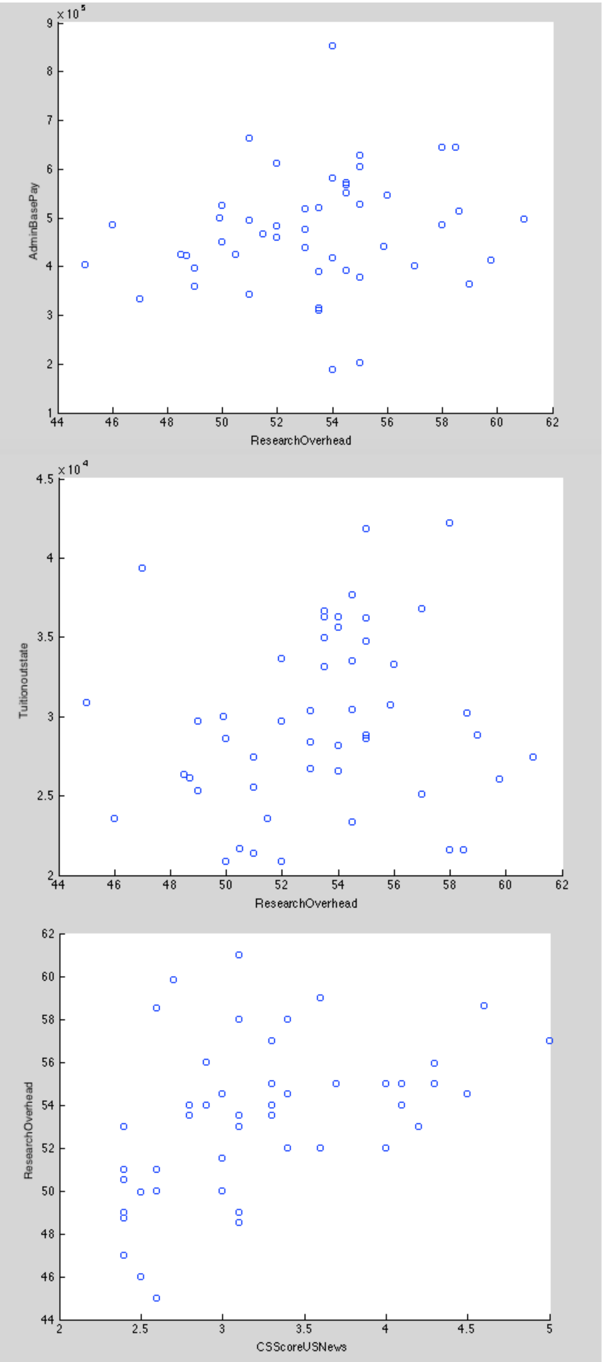
The given dataset is a subset of the top 100 US universities containing five columns representing Score given by US News, Research overhead, Admin base pay, Tuition cost for out-of-state students and number of students enrolled. We consider every column as a random variable and continue with our mathematical approach.

3. METHODS

We initially calculate mean, variance, standard deviation for each of the four variables considered, the covariance and correlation matrices and the log likelihood of the data assuming that each of the variables is normally distributed and that they are independent of each other.

Scatter Plots of the pairwise data shown below throw some light on the correlation between the variables. From observation CS Score and Research Overhead are most related whereas Tuition and Admin Base Pay are least correlated.



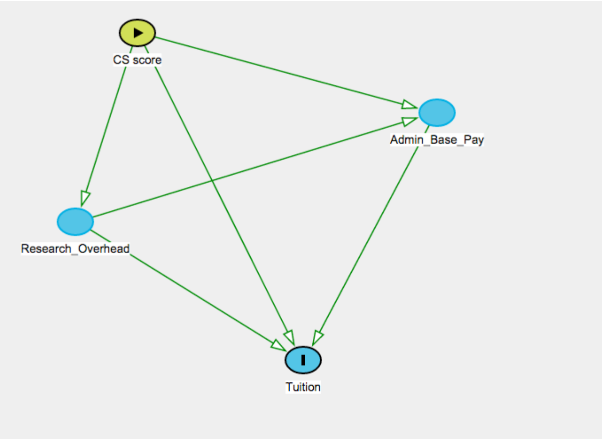


Using normlike function in Matlab we calculated the log-likelihood for the dataset which came out to be -1314.66855.

4. BAYESIAN CONSTRUCTION

After knowing the maximum likelihood estimation from other calculations we are targeting to achieve an optimal likelihood higher than the previous one. A Bayesian network would be constructed to prove this is possible.

We have created a Bayesian Network based linking the different variables of the directed acyclic graph on observing the correlation matrix values among different variables and also taking into consideration the causal effects of some variables on the other given we have some domain specific knowledge. Two nodes are connected directly if one affects or causes the other. We see CS Score and Research Overhead have strong correlation also it can be reasoned based on knowledge that CS score influences other variables like Research Overhead, Tuition and Admin Base Pay thus we assume CS score as parent of all other. Research Overhead causes Admin Base pay to go up also influences tuition which is captured by making Research Overhead as parent of tuition and Admin Base pay. Also Admin Base pay may cause the tuition to go up which can be captured by making Admin Base pay as parent of tuition. Based on our causal reasoning the Bayesian Graph is given below.



For finding the likelihood, we calculate the joint probability of the Bayesian network using the formula given below.

$$P(x_1,x_2,\ldots,x_n) = \prod_i P(x_i|Parents(X_i))$$