Probability Distributions and Bayesian Networks

Mihir Kulkarni
Computer Science and Engineering
University at Buffalo
Buffalo, NY-14214
UB Person number:50168610
mihirdha@buffalo.edu

ABSTRACT

This project calculates probability distributions and various statistical properties like mean, variance, standard deviation and log likelihood for a data set of American universities. Using the covariance and correlations between the attributes of the universities, it constructs a Bayesian network representing the dependencies between the attributes. The project also finds out the optimal Bayesian network by calculating the log likelihood.

***Index Terms*—Bayesian network, log likelihood, mean, variance, standard deviation.**

I. Introduction

The study of machine learning is highly dependent upon various probability distributions and other statistical properties. Hence, calculating these properties is essential to t the study of Machine learning. We are calculating here – mean, variance, log likelihood, covariance matrix, and correlation matrix. Using these properties we calculate optimal Bayesian network for the given attributes of the data set. Optimal Bayesian network is found out by selecting the Bayesian network with highest log likelihood. We can also find some interesting conditional probabilities and dependencies using optimal Bayesian network.

We are implementing this project on MATLAB software by Mathworks.

II. Data set

The data set is the data of 50 American universities. Each university has attributes namely – rank, name, CS score, research overhead, admin base pay, out of state tuition and CS grad student number. For this project we are only focusing on four of these attributes- CS score, research overhead, admin base pay and out of state tuition. These attributes are labelled X1, X2, X3, and X4. The fifth attribute X5 i.e. CS grad student number has missing values. This data has been provided for the project in excel spreadsheet titles UniversityData.xls.

III. Mean, variance and standard deviation

Arithmetic mean of the data is calculated using the formula-

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Where, X_1, X_2… Xn are the values of an attribute of the data set. In MATLAB, we have calculated the mean using the MATLAB function 'mean ()'.

Variance of the data set is given by

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

In the project, we are calculating the variance by using the MATLAB function 'var ()'.

Standard deviation is calculated by finding out the square root of the variance. In the project, we are calculating the standard deviation by using the MATLAB function 'std ()'.

IV. Covariance and correlation

Covariance and correlation are measures of how the two attributes are related to each other. In this project we have found covariance and correlation of each pair of variables. Using these values, we have created covariance matrix and correlation matrix. These matrices are useful in determining the internal relationships of the data set. These matrices are found using MATLAB functions 'cov ()' and 'cor ()' respectively.

Larger value of correlation signifies greater correlation. Among the data set, **most correlated variables are CS Score and Research overhead**. **Least correlated variables are Admin base pay and tuition fee** of the university.

V. Log-likelihood

We are calculating the log-likelihoods of each attribute separately assuming that each attribute is normally distributed and is independent of other attributes. This is nothing but the probability for each attribute. We get the log likelihood of our data set after adding these four individual values.

$$\text{logLikelihood} = \sum \text{sum}(\log(\text{normpdf}(\text{Variable}_i, \text{mean}_i, \text{std}_i)))$$

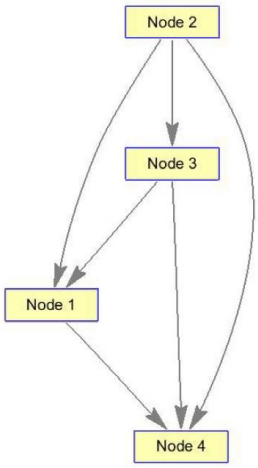
VI. Bayesian network

We have found the optimal Bayesian network by performing exhaustive search on all possible networks of the data. We first find whether the candidate is a directed acyclic

graph. Then we calculate the log likelihood of the candidate. We select the Bayesian network with highest log-likelihood.

The optimal Bayesian network we have got is

0	0	0	1
1	0	1	1
1	0	0	1
0	0	0	0



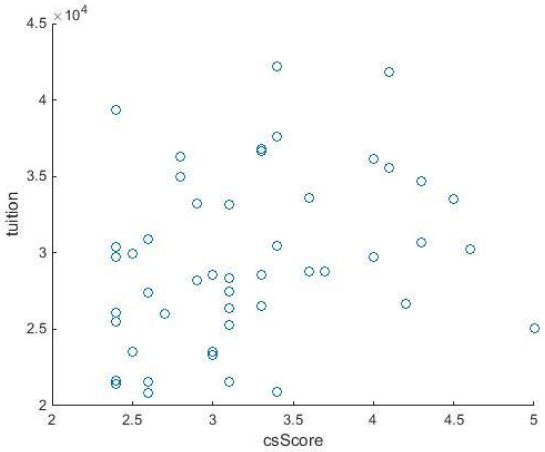
Log-likelihood of this network is -1304.1

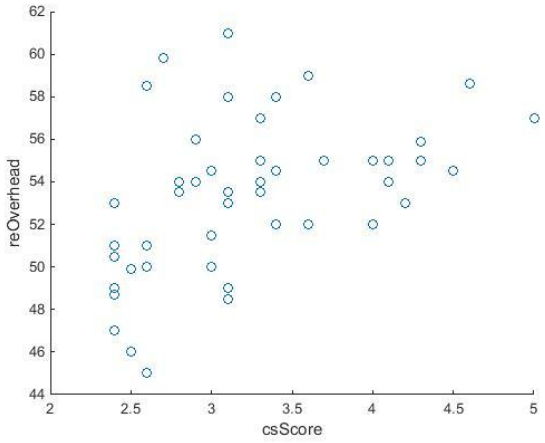
We can see from this Bayesian network that

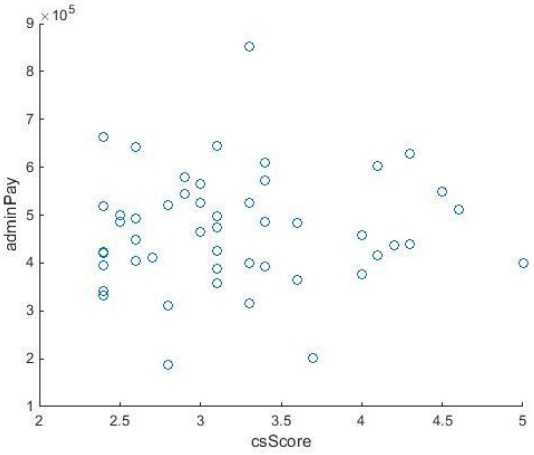
- **Tuition is dependent on all three variables.**
- **CS score is dependent on research overhead and admin base pay.**
- **Admin base pay is dependent on research overhead.**
- **Research overhead is independent of all variables.**

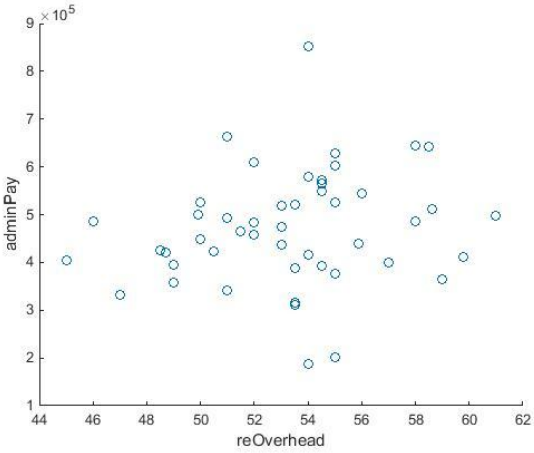
VII. PLOTTING VARIABLES

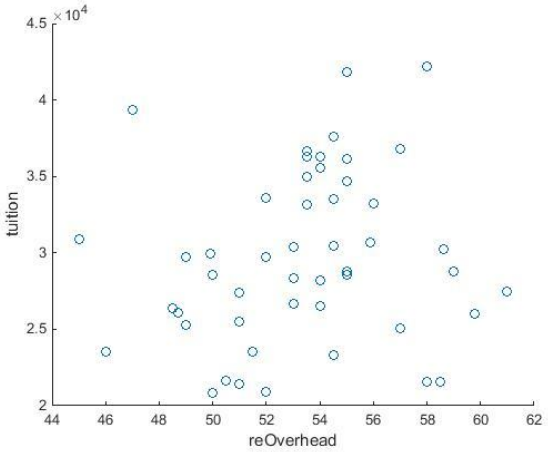
We are plotting all the pairs of variables against each other using scatter and text commands of MATLAB

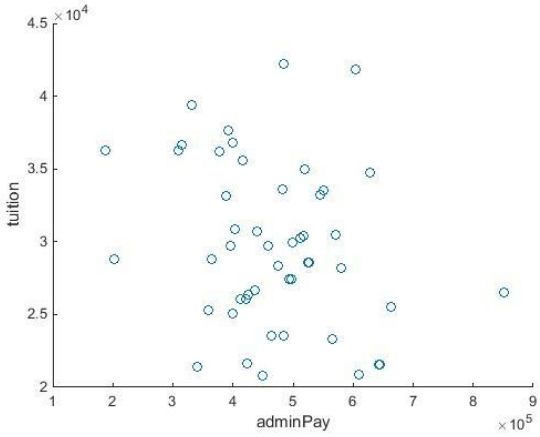












REFERENCES

[1] Wikipedia page on Mean,Variance and Standard deviation.

[2] “http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html”

[3] Class material and TA’s guidance.