Probability Distributions & Bayesian Networks

Raghavendra Prakash Nayak

***Abstract*—Probability distribution for different variables is the main focus of this project. The project involves the learning MATLAB to evaluate sufficient statistics: mean, variance and standard deviation of univariate distributions and covariance and correlation coefficient of variables. These will be used to construct joint probabilities or Bayesian Networks. The project also calls for evaluation of these representations by using likelihood.**

***Index Terms*— Bayes Methods, Correlation coefficient, Covariance matrices, Directed Acyclic Graphs, Probability Distribution, Maximum Likelihood estimation.**

.

I. Introduction

BAYESIAN Networks are a widely used for capturing probabilistic relationships among variables. They can be used to provide a compact joint probability distribution by capturing the dependency among the variables. The joint probabilities or their conditional dependencies can be represented by Directed Acyclic Graphs (DAG). These DAGs represent random variables in a Bayesian sense. The edges represent conditional dependencies and the variables or nodes that are not connected are termed conditionally independent.

For many applications, the log-likelihood is more convenient to work with as it is a monotonically increasing function. It achieves its maximum value at the same point of the function itself and can hence be used in place of the maximum likelihood estimation. In a situation where the parameters explain a collection of statistically independent observations, the likelihood function factors into a product of individual likelihood functions and the logarithms of this product is a sum of individual logarithms.

In this project we focus on a dataset of university rankings, tuition etc… based on a survey conducted by US News and Chronical to construct a Bayesian network with high log-likelihood. Furthermore we try to use these Bayesian networks to determine some conditional probabilities.

II. The Dataset

For this project, I used the dataset made available by US News for the ranking of colleges for MS in Computer Science.

The data set consists of 49 of the top 100 public universities in the US, graded on the computer science score, the research overhead, the admin base pay, tuition fees and the number of students enrolled in the MS in computer science course.

The X1 (CS ranking score) corresponds to the score of a public university according to a survey conducted by US News. These are a subset of the top 100 Computer Science graduate programs in the US according to US News and World Report. X2(Research Overhead) corresponds to the portion of research grants retained as administrative/infrastructure costs by the university. X3 (Administrator Base Salary in $) corresponds to the average salary of the administrator as indicated by Chronical. X4(Tuition-out of state in $) corresponds to the tuition for the computer science graduate program for out-of-state students. And X4(No of graduate students in Fall2015) corresponds to the number of graduate students enrolled for the computer science program in these colleges for the fall 2015 term. The scatter plots of the variables of the data set is given below:

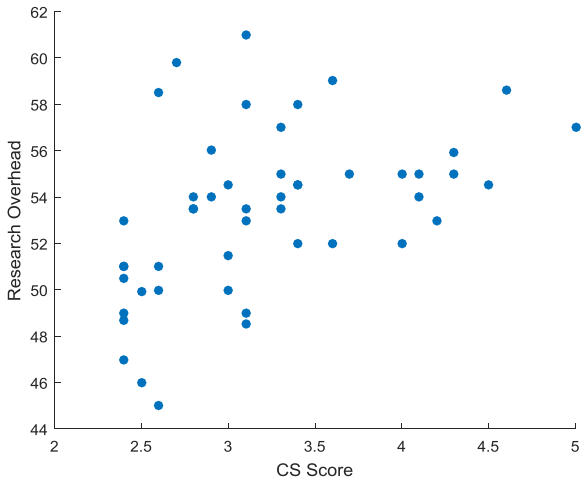


Figure-1: Scatter plot of CS Score vs. Research Overhear

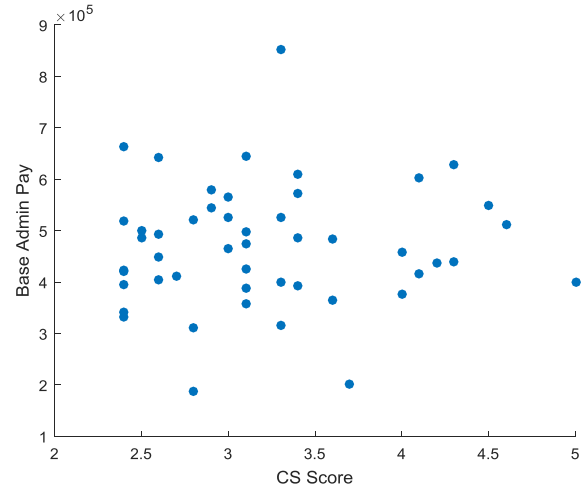


Figure-2: Scatter plot of CS Score vs. Admin Base Pay

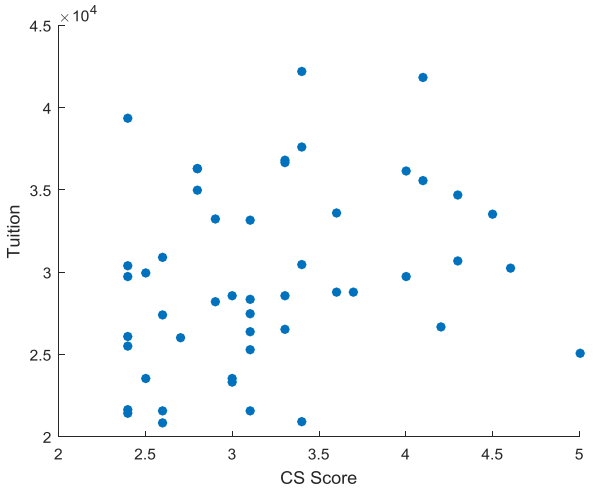


Figure-3: Scatter plot of CS Score vs. Tuition

III. THE METHOD

To construct the Bayesian network with a high log likelihood, this project is divided into different tasks with each task depending on the completion of the prior. The columns X1, X2, X3 and X4 are considered in constructing the Bayesian network.

A. Computing the Mean, Variance and Standard Deviation

This involved in computing for each variable (X1, X2, X3 and X4) the mean, variance and the standard deviation. The values of these were stored in variables mu1, mu2, mu3, mu4, var1, var2, var3, var4, sigma1, sigma2, sigma3 and sigma4. Where, mu, var and sigma correspond to mean, variance and standard deviation respectively.

The mean of the variables was computed by using the formula;

$$\text{mu} = \mu = \sum_{i=1}^{N} x(i)$$

The corresponding MATLAB function used was; *mean (X)*

The variance of the variables was computed by using the formula;

$$\text{var} = \sigma^2 = 1/(N-1) \cdot \sum_{i=1}^{N} [x(i) - \mu]^2$$

The corresponding MATLAB function used was; *var (X)*

The standard deviation of the variables was computed by using the formula;

$$\text{sigma} = \sigma = \sqrt{\text{var}} = \sqrt{[1/(N-1) \cdot \sum_{i=1}^{N} [x(i) - \mu]^2]}$$

The corresponding MATLAB function used was; *std(X)*

B. Computing Covariance and Correlation Matrices

We then computed the covariance and the correlation for the variables of the dataset. The covariance and correlation were in the form of a 4x4 matrix for the given dataset.

The covariance was computed for the dataset (X1, X2, X3 and X4) using.

$$\text{Covariance} = \sigma_{123\ldots d}$$

$$= 1/(N-1) \sum_{i=1}^{N} [x_1(i) - \mu_1]\,[x_2(i) - \mu_1] \ldots [x_d(i) - \mu_1]$$

The corresponding MATLAB function used was; *conv(X)*

Correlation Coefficient was computed by;

$$\rho = \frac{\sigma_{123\ldots d}}{\sigma_1 \sigma_2 \ldots \sigma_d}$$

The corresponding MATLAB function used; *corrcoef(X)*

C. Computing Log-Likelihood

This involved computing a Gaussian (normal) distribution of the variables and then using the normalized values we computed the log-likelihood of the dataset.

The normal distribution was computed using the formula;

$$p(x) = (1/\sqrt{2\,\Pi\sigma})\, \exp[\, -1/2 \,.\, ((x - \mu)/\, \sigma)^2]$$

The corresponding MATLAB function used was; *normpdf(X)*

Once the normal distribution was found for all the variables of the data set, a logarithm of the variable set was taken using the *log* function in MATLAB.

Ex: log (normpdf(X1))

Once the logarithm function was completed the sum of the 1x49 matrix of X1, X2, X3 and X4 was taken. And finally to get the log likelihood the logs of X1, X2, X3 and X4 were added together.

Ex:

```
logLikelihood = (sum(log(normpdf(X1)))) +
(sum(log(normpdf(X2)))) +  (sum(log(normpdf(X3)))) +
(sum(log(normpdf(X4))))
```

D. Computing Bayesian Network Log-Likelihood and Bayesian Network Graph

In order to find the Bayesian Network Log-Likelihood the Directed Acyclic Graphs for the maximum possibilities of a 4x4 matrix were first found. For this, all the possible values of a 4x4 matrix were considered which were equal to 2^{16} in binary equal to 65535 possible graphs. From these the DAG were separated out using the *graphisdag()* function.

Ex.

```
For x= 1:65535
    g = dec2bin(x,16)    %to make it 16 bits long
    r =reshape(str2num(reshape(g()',[],1)',[4,4])
    DAG=[];
    if graphisdag(r)
        DAG(:,:,i)=r
```

Once the DAGs were obtained, these were computed with the dataset that was given to find the parent and child nodes in each of the rows. Using the edges that were defined in the DAGs, the multivariate probabiltiy distribution was computed. Using this, the highest value of the Bayesian Network Log-Likelihood was found and the Bayesian Network graph was obtained. The formula used to compute the multivariate cumulative distribution used was;

$$F(x1,x2) = P(X1 \leq x1, X2 \leq x2) = \int\int p(x1,x2).dx$$

The corresponding MATLAB function for multivariate probability distribution used was: *mvnpdf(X,mean(X),cov(X))*

IV. RESULTS

From the dataset, using MATLAB functions and incorporating the steps as delineated in the tasks, the results for each of the project tasks were found.

The result for the first task of computing the mean, variance and standard deviation into variables mu1, mu2, mu3, mu4, var1, var2, var3, var4, sigma1, sigma2, sigma3 and sigm4 for X1, X2, X3, X4 of the dataset respectively are as below:

Variable	Value	Variable	Value
mu1	3.2142857142857	mu3	469178.81632653
mu2	53.365306122448	mu4	29711.959183673
var1	0.4574999999999	var3	1.4189720820903
var2	12.616062925170	var4	3.1367695789965
sigma1	0.6763874629234	sigma3	119120.61459253
sigma2	3.5519097574642	sigma4	5600.6870819539

From the dataset, the second task of computing the 4x4 matrices for covariance and correlation was completed. The result is below;

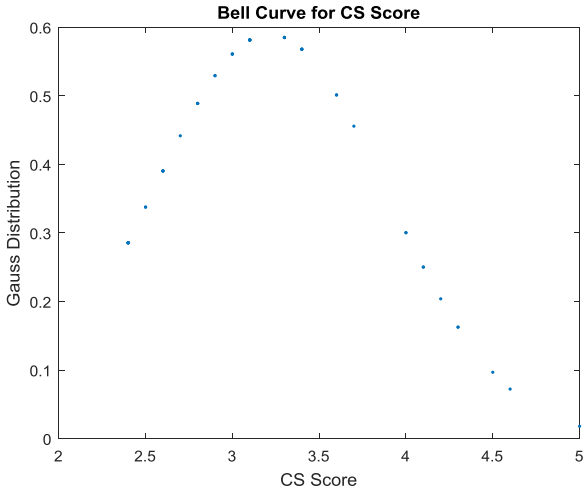
Covariance matrix:

0.4575	1.1184	3879.78	1058.47
1.1184	12.616	66651.6	2975.82
3879.78	66651.6	1.41897	-1.6368
1058.47	2975.82	-1.6368	3.13676

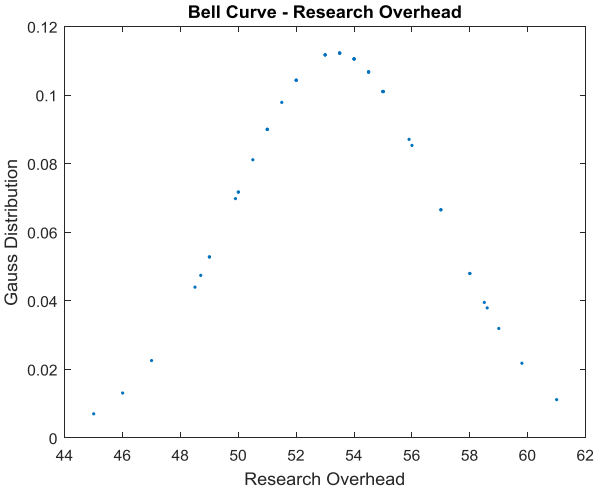
Correlation matrix:

1.0	0.4655	0.0481	0.2794
0.4655	1.0	0.1575	0.1495
0.0481	0.1575	1.0	-0.2453
0.2794	0.1495	0.2453	1.0

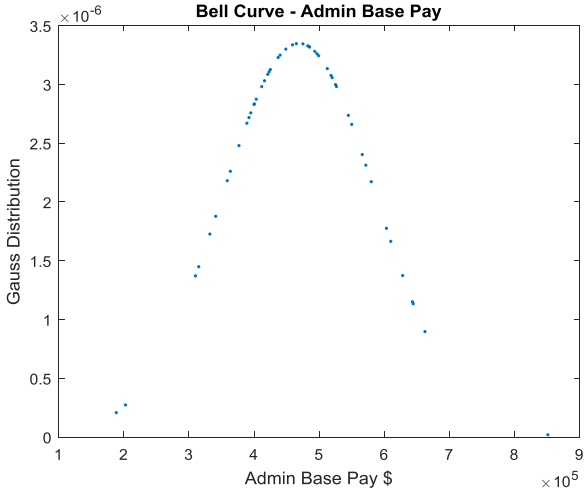
The Gauss Distribution plot for X1(CS Score):



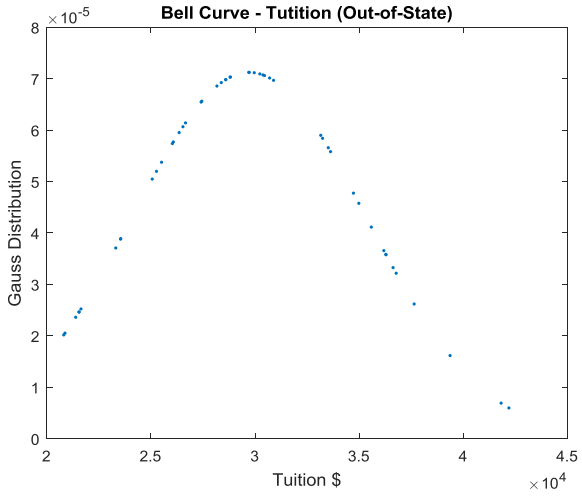
The Gauss Distribution plot for X2 (Research Overhead):



The Gauss Distribution plot for X3(Admin Base Pay):



The Gauss Distribution plot for X4(Tuition):



The result of the Log-Likelihood for the given dataset was found to be: -1314.66855043451

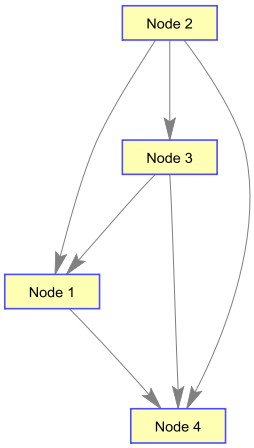
Post finding the result, for the dataset of 4 variables, X1, X2, X3 and X4, the maximum possible graphs were found to be: Maximum possible joint probability graphs = 2^{16} which is 65535. For these, the maximum possible Directed Acyclic Graphs were found to be 532.

Using this, the BNLogLikelihood was calculated to be: BNlogLikelihood = -1304.09236999962.

BNGraph =

$$\begin{vmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

Bayesian Network Graph Plot:



V. CONCLUSION

The mean, variance and standard deviation for the given dataset has been successfully computed and these have been used to find the log-Likelihood of the dataset. The covariance and correlation matrices for the dataset has also been found. Using the covariance and Directed Acyclic Graphs, the joint probability or Bayesian Network has been found with the highest log-Likelihood, the graph of the same has also been found.

VI. REFERENCES

[1]. Report Format – "From Classical to Hip Hop: Can Machines Learn Genres" Aron Karvitz, Eliza Lupone, Ryan Diaz–(https://d1b10bmlvqabco.cloudfront.net/attach/idx2of0x5ho7hk/hbze98s14z32mo/iew91o0mzm8g/Aaron_Kravitz_Eliza_Lupone_Ryan_Diaz_Can_Machines_Learn_Genres.pdf).

[2]. "Bayesian Networks" – Wikipedia (https://en.wikipedia.org /wiki/Bayesian_network).

[3]. "Likelihood Function" – Wikipedia (https://en.wikipedia.org/wiki/Likelihood_function).