Project 1

Probability Distributions and Bayesian Networks

CSE 574
Introduction to Machine Learning

Anirudh Reddy Nalamada
UB No.: 5016-9240
UBIT Name: aniredn

1. Introduction

The main aim of this project is to understand the basic functionality of MATLAB and demonstrate and understanding of the basic concepts of probability such as mean, variance, correlation, likelihood and Bayesian Networks.
The data set given contains four variables CS Score, Research Overhead, Admin Base Pay and Tuition. Each variable has 49 entries corresponding to 49 different colleges.

2. MATLAB Functions Used

i. xlsread(*filename*) – Reads the data from a file of format xls
ii. mean(X) – Calculates the mean of elements of matrix/array X
iii. var(X) – Calculates the variance of elements of matrix/array X
iv. std(X) - Calculates the standard deviation of elements of matrix/array X
v. cov(X) – Calculates the covariance of elements of matrix X. Returns the value in the form of a matrix
vi. corrcoef(X) – Calculates the correlation coefficient of matrix X. Returns the value in the form of a matrix
vii. plotmatrix(X) – Creates a scatter plot from the data in Matrix X.
viii. normpdf(X) – Calculates the normal probability distribution function of dataset X.
ix. mvnpdf(X) – Calculates the multivariate normal probability distribution function of dataset X.
x. sum(X) – Calculates the sum of elements present in matrix/array X.
xi. log(X) – Calculates the logarithmic value of the element X.

3. Discussion of Covariance and Correlation of Variables

The covariance and correlation data was calculated from the given data set. Covariance and Correlation values are represented as a 4x4 matrix.

Table 1: Covariance matrix of the dataset

X	CS Score (CSS)	Research Overhead (RO)	Admin Base Pay (ABP)	Tuition (T)
CS Score (CSS)	0.4575	1.118422619	3879.781845	1058.479762
Research Overhead (RO)	1.118422619	12.61606293	66651.66433	2975.829804
Admin Base Pay (ABP)	3879.781845	66651.66433	14189720821	-163685641.3
Tuition (T)	1058.479762	2975.829804	-163685641.3	31367695.79

Table 2: Correlation matrix of the dataset

X	CS Score (CSS)	Research Overhead (RO)	Admin Base Pay (ABP)	Tuition (T)
CS Score (CSS)	1	0.465530853	0.048153164	0.279412416
Research Overhead (RO)	0.465530853	1	0.157529593	0.149590791
Admin Base Pay (ABP)	0.048153164	0.157529593	1	-0.245347903
Tuition (T)	0.279412416	0.149590791	-0.245347903	1

It can be observed from Table 2 that CS Score (CSS) and Research Overhead (RO) are more strongly correlated than any other pair of variables followed by CSS and T, ABP and T, RO and ABP, RO and T and the least being CSS and ABP.
Figure 1 clearly displays the how the values of variables change with respect to each other.

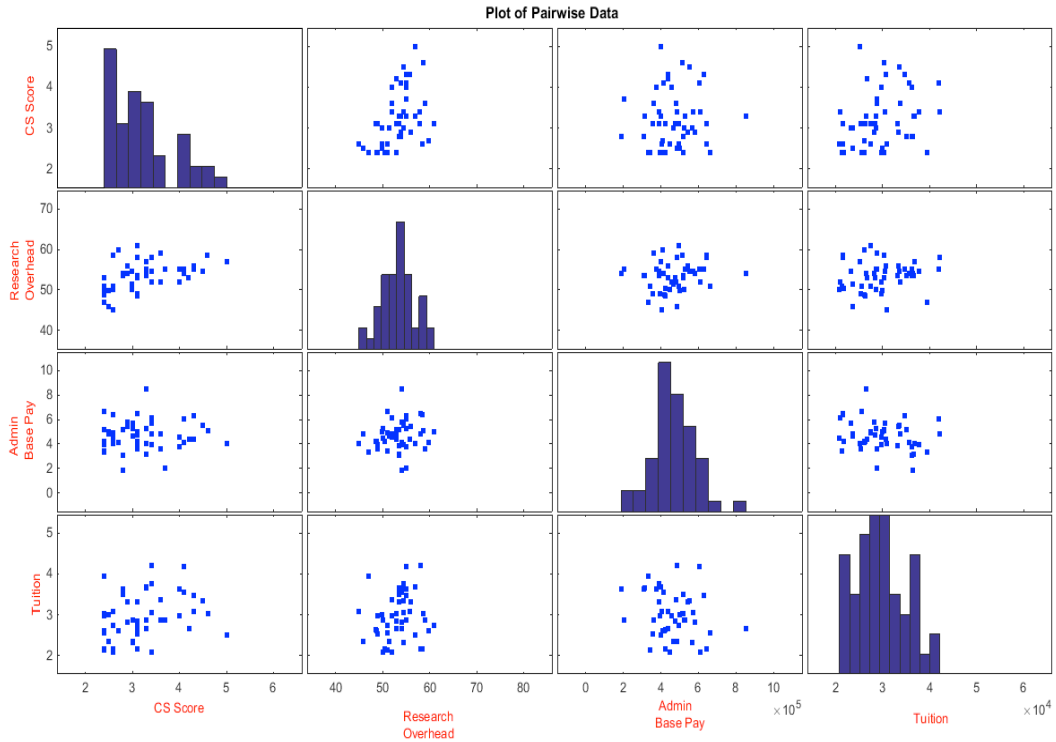

Figure 1: Plot of Pairwise Data

4. Determination of the Bayesian Network

Using the correlation values the following Bayesian Network has been constructed.

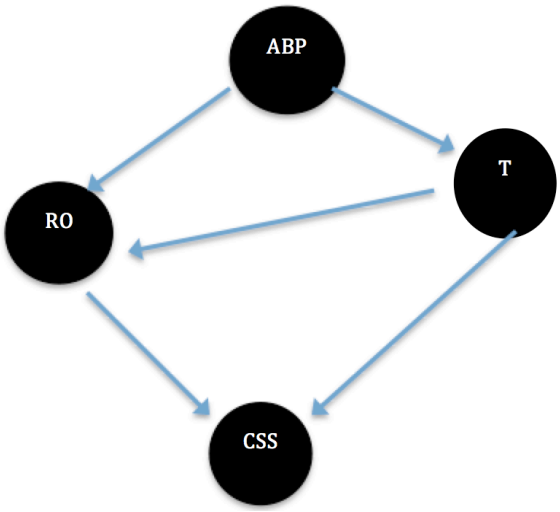

Figure 2: Plot of Pairwise Data

- In the given data set, CS Score was the basis on which the colleges were ranked. Hence it can be assumed that it is dependent on all other variables and no variable is dependent on the CS Score. Using the correlation values, it can be determined that the Research Overhead and the Tuition fee strongly influence the CS Score. The correlation value for Admin Base Pay and CS Score is almost negligible hence it has been ignored.
- There is a negative but a significant correlation between Tuition and Admin Base Pay. Hence a link has been established between Tuition and Admin Base Pay. Since it is a negative correlation the value of Tuition increases when Admin Base Pay decreases.
- Research Overhead value is equally dependent on Admin Base Pay and Tuition as observed from the correlation values.

The above Bayesian Network can be represented as the following matrix, which was stored in the BNgraph variable.

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

For this Bayesian Network, the probability function can be expressed as below:

$$P(CSS, RO, ABP, T) = P(CSS \mid RO, T) * P(RO \mid ABP, T) * P(ABP) * P(T \mid VBP)$$

The Log Likelihood value of the data for the Bayesian Network was found out to be approximately -1.3041e+03 as compared to -1.3146e+03 originally.

5. Conclusion

The Bayesian Network constructed in this project has been created purely based on a logical analysis of the data. Furthermore, an algorithm can be implemented to calculate the optimal Bayesian Network for any given data set.

6. References

- Daphne Koller and Nir Friedman, "Probabilistic graphical models: principles and techniques", MIT Press 2009
- Christopher M. Bishop, “Pattern Recognition and Machine Learning”, Springer 2006