Project Report: Probability Distributions and Bayesian Networks

CSE 574 – Fall 2015
Aniket Thigale
50168090

1 Problem Introduction

Machine Learning methods are based on probability theory and statistics. This project concerns probability distributions of several variables. We will learn how to use MATLAB to evaluate sufficient statistics: mean and variance of univariate distributions and covariance and correlation coefficient of pairs of variables. We will then use these statistics to construct compact representations of joint probability distributions known as Bayesian networks. Then we will evaluate the goodness of these representations by using the concept of likelihood. Finally we will use the Bayesian networks to answer some queries.

2 Dataset

The multivariate data, present in the form of an excel sheet UniversityData.xls, has four variables, namely, CS Ranking Score, Research Overhead, Admin Base Pay and Tuition along with a fifth variable CS Grad Student No. which has some missing values. This data has been obtained from various sources like US News and World Report, various university websites and chronicle. Each value corresponds to the score of a public university according to a survey conducted by US News and World Report. Research Overhead corresponds to the portion of research grants retained as infrastructure/administrative costs by the university.

3 Useful Terms

Mean, Variance and Standard Deviation:

The sample mean of a univariate distribution of variable X with N samples x(i); i = 1.. N has the form:

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

Variance measures how far a set of numbers is spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data points tend to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data points are very spread out around the mean and from each other.
The sample variance is computed as

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} [x(i) - \mu]^2$$

Where σ is referred to as the standard deviation.

Covariance:

In probability theory and statistics, covariance is a measure of how much two random variables change together.

Correlation Coefficient:

A correlation coefficient is a coefficient that illustrates a quantitative measure of some type of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values.

Univariate Distribution:

A univariate distribution is a probability distribution of only one random variable.

Multivariate Distribution / Joint Probability Distribution:

The joint probability distribution for any no of random variables X, Y,… is a probability distribution that gives the probability that each of X, Y, ... falls in any particular range or discrete set of values specified for that variable.

Probability Mass Function:

A probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

Probability Density Function:

A probability density function (PDF), or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value. The probability of the random variable falling within a particular range of values is given by the integral of this variable's density over that range.

Log Likelihood:

A likelihood function is a function of the parameters of a statistical model. It is used when describing a function of a parameter given an outcome.

Given N independent samples x1,… xN from a probability distribution p(x), the log-likelihood of observing the samples is given by:

$$\mathbf{L}(\mathbf{x}_1, ..\mathbf{x}_N) = \sum_{i=1}^{N} log\ p(\mathbf{x}_i)$$

Conditional Probabilities:

Given two variables X1 and X2 the sum rule of probability is:

$$p(x_1) = \sum_{Val(x_2)} p(x_1, x_2)$$

Where V al is the set of values taken by its argument. The sum rule allows us to obtain the marginal probability p(x1) from the joint probability p(x1; x2). The product rule of probability is:

$$p(x_1, x_2) = p(x_1|x_2)p(x_2)$$

From which we get the chain rule:

$$p(x_1, x_2, x_3) = p(x_1|x_2, x_3)p(x_2|x_3)p(x_3)$$

Bayesian Network Factorization:

Given a Bayesian network *G* of *N* variables **X** = {X1,.. Xd}, the joint probability distribution is given by

$$p(\mathbf{X}) = \prod_{i=1}^{N} p(X_i|pa(X_i))$$

where pa(Xi) are the parent variables of Xi.

4 APPROACH

1. Calculate the mean, variance and standard deviation for each variable
2. Calculate the covariance and correlation between any 2 pairs of variables
3. Construct the final 4x4 covariance and correlation matrices
4. Plot the graphs for data points between any 2 variables
5. Now there are two ways to calculate the logLikelihood :
   - 5.1. Calculate the logLikelihood for each variable using normpdf and sum them
   - 5.2. Calculate the logLikelihood for all variables using mvnpdf
6. Constructing the optimum Bayesian Network resulting in higher logLikelihood than in (5)
   - 6.1. We are using the exhaustive search to find the optimum Bayesian Network. There are 65536 possibilities which we store in a 65536x16 matrix
   - 6.2. Next using each possibility, we check if it is cyclic using graphisdag() matlab function
   - 6.3. If it is a DAG, then for each column Find if ones exist in the column. The row nos. are the parents
   - 6.4. If no ones exist, calculate and add the logLikelihood for this column using normpdf()
   - 6.5. Else calculate and add the logLikelihood using mvnpdf() and Bayesian Network Factorization formula. Calculate the denominator based on
   - 6.6. Compare and store the highest logLikelihood and Bayesian Network graph.
   - 6.7. Using biograph() display the graph

5 RESULTS

1. Looking at the university data, CS Ranking score and Research overhead are the most correlated variables while Administrator Base pay and CS Ranking are the least correlated variables.
2. Optimum Bayesian Network obtained:

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

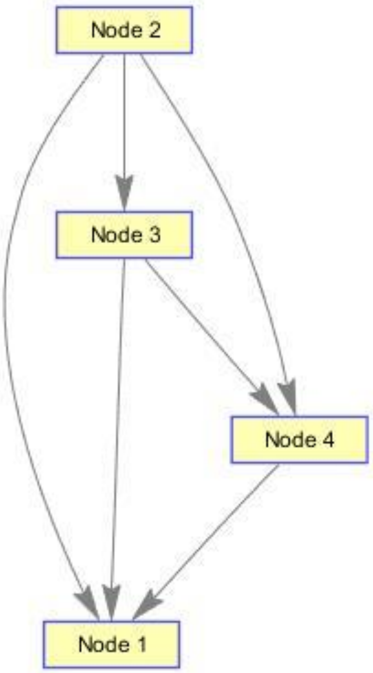

Fig: Optimum BN Graph

Looking at the graph we can say that Research Overhead is an independent variable. Administrator Base Pay is dependent on research overhead. Tuition is dependent on Research Overhead and Administrator Base Pay. CS Ranking Score is dependent on all other variables.

6 REFERENCES

1. Probabilistic Graphical Models by Daphne Koller
2. Class notes
3. www.wikipedia.com
4. www.mathportal.org