Project 1

Probability Distributions and Bayesian Networks

CSE474/574: Introduction to Machine Learning (Fall 2016)

Instructor: Sargur N. Srihari

Teaching Assistants: Jun Chu, Junfei Wang and Kyung Won Lee

Sugosh Nagavara Ravindra

sugoshna@buffalo.edu

Person#: 50207357

Table of Contents

Abstract

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. In this project we try to construct a Bayesian network for four variables (CS Score, Research Overhead, Admin Base Pay, Tuition). In order to construct a Bayesian Network we need to map it to the correlation values of the variables. Initially we perform few statistical calculations on the data and calculate the log likelihood. The main aim of this project is to maximize the log likelihood using the Bayesian graph and conditional probabilities.

Considering the variables to be independent we obtained a log-likelihood of **-1315.119.** After constructing Bayesian Network and learning its parameters we reduced the log-likelihood to **-1305.880.** The Bayesian network can be constructed by studying the correlation values and mapping highly correlated values to each other.

Introduction

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine Learning algorithms focus on learning how the data is modeled and subsequently developing a prediction system for future data inputs. Design of learning algorithms often relies on probabilistic assumption of the data. In this report we will look into one of the machine learning algorithms revolving around probability theory named Bayesian networks.

We have university data for 49 different colleges and the data comprises of CS Score, Research Overhead, Admin Base Pay, Tuition. Our aim of this project maximize the log-likelihood by constructing a Bayesian Network comprising of variables which are highly correlated.

Statistical Concepts

3.1 Mean: Mean of a continuous variable is the sum of all its values divided by total number of values.

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

3.2 Variance: Variance informally measures how far a set of (random) numbers are spread out from their mean

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

σ refers to standard deviation.

3.3 Covariance: Covariance is the measure of how pair of variables are related to each other. A positive value indicates that the variables are positively related, meaning if one increases the other increase too. A negative covariance indicate the variables are negatively related. The values of covariance does not mean anything. It's just the sign we need to look for.

$$\sigma_{12} = \frac{1}{N-1}\sum_{i=1}^{N}[x_1(i) - \mu_1][x_2(i) - \mu_2]$$

Following is the covariance matrix obtained for the university data.

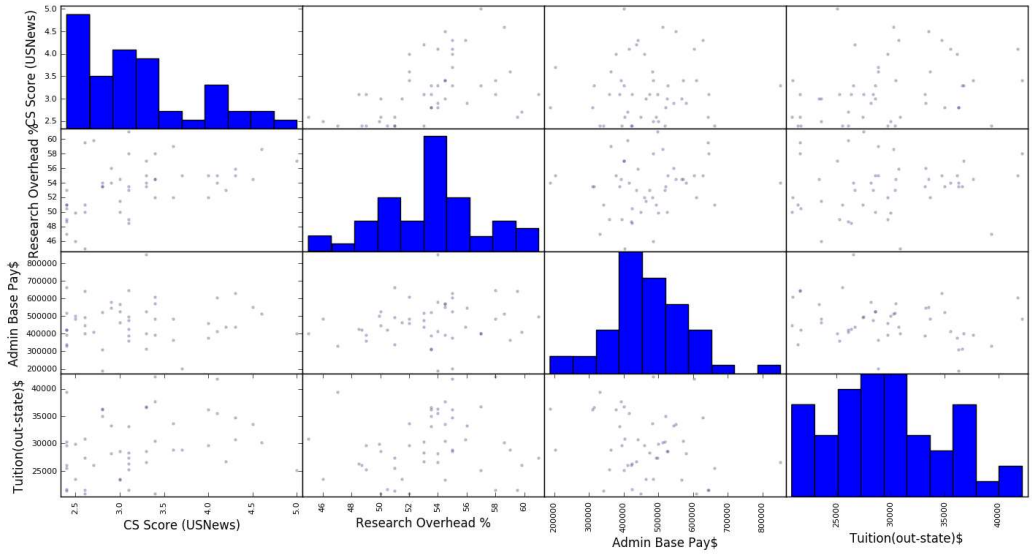
	CS Score	Research Overhead	Admin Base Pay $	Tuition $
CS Score	0.458	1.106	3879.782	1058.48
Research Overhead	1.106	12.85	70279.376	2805.789
Admin Base Pay $	3879.782	70279.376	14189720820.903	-163685641.258
Tuition $	1058.48	2805.789	-163685641.258	14189720820.903

3.4 Correlation: Correlation is the measure of how the pair of variables are related and represented with and actual value (scale) of how much they are related.

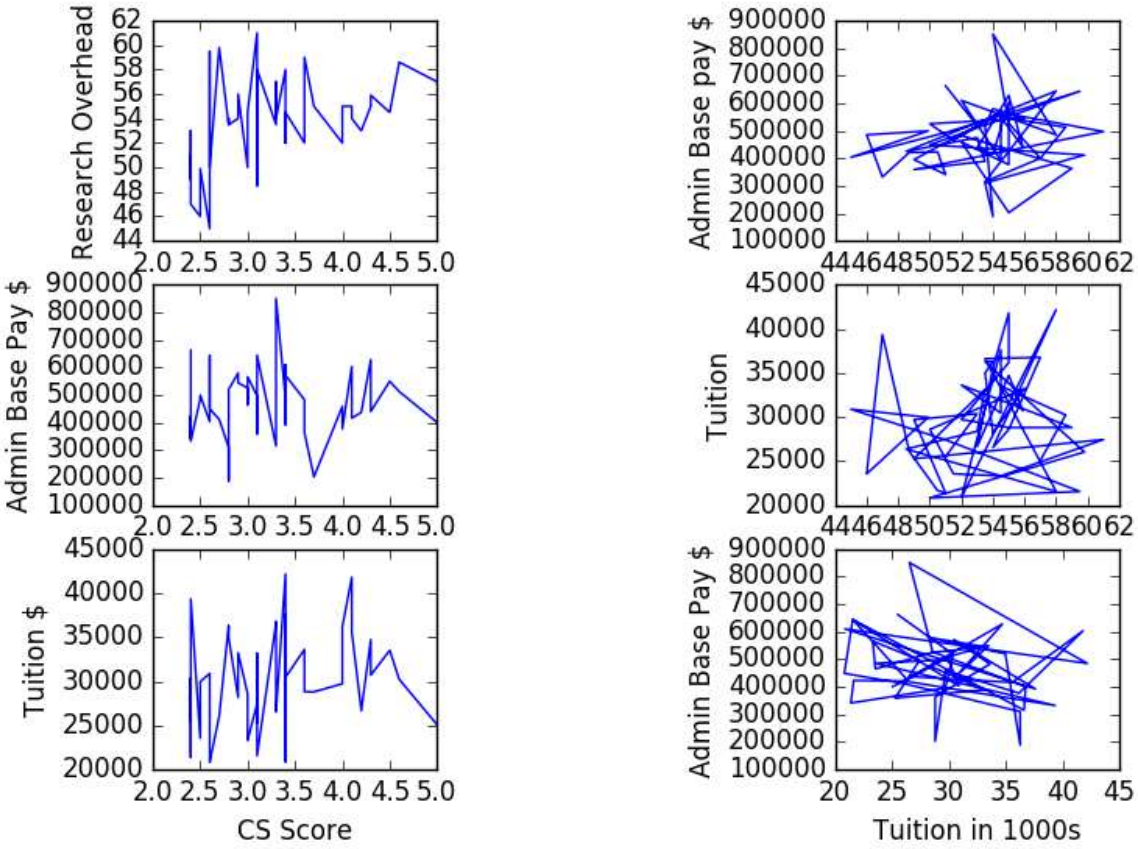
Following is the correlation matrix obtained for the university data

	CS Score	Research Overhead	Admin Base Pay $	Tuition $
CS Score	1.0	0.456	0.048	0.279
Research Overhead	0.456	1.0	0.165	0.14
Admin Base Pay $	0.048	0.165	1.0	-0.245
Tuition $	0.279	0.14	-0.245	1.0

Following is the graph obtained for pairwise correlation data



From graph and correlation matrix we can infer that CS Score and Research overhead are the most correlated and CS Score and Admin Base Pay are the lease correlated.



The above figure represents pair wise relationship between raw data values.

3.5 Log-Likelihood: Likelihood is used after data are available to describe a function of a parameter for a given outcome. Log-likelihood is the sum over log values of all probabilities derived from the probability distribution function over all the variables
Following are the log-likelihood values of the four values from university data:

- CS Score: -49.869
- Research Overhead: -131.587
- Admin Base Pay: -641.734
- Tuition: -491.929

Summing all of the above values will get us the Log-Likelihood if we consider all the variables as independent. Log-Likelihood: **-1315.119**

Bayesian Network

A Bayesian network or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph. Once we construct the Bayesian network by studying the correlation matrix we factorize it by forming set of joint and conditional probabilities. The joint probability distribution is calculated using the formula

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

We then find the Log-Likelihood using

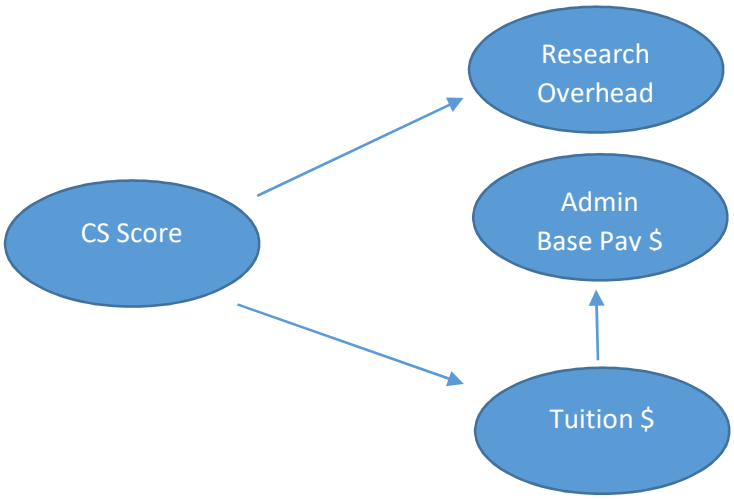
$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}[1], ..\mathbf{x}[N]) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\beta_0 + \beta_1 x_1[n] + ..\beta_k x_k[n] - y[n])^2$$

We can derive Beta and Sigma values by partially differentiating and finally find the Log-Likelihood. Our main aim is to maximize the Log-Likelihood. We can construct the Bayesian network by studying the correlation matrix. We see that:

- CS Score and Research overhead are highly correlated.
- CS Score and Tuition are positively correlated.
- Tuition and Admin Base pay is inversely correlated, meaning if one increases the other decreases

From the above observation we can construct a Bayesian Network denoting all relations.

Following is a Bayesian network obtained from the university data.



Log(P(CS_Score)): -49.870

Log(P(ResearchOverhead|CS_Score)): -125.870

Log(P(AdminBasePay|Tuition)): -640.208

Log(P(Tuition|CS_Score)): -489.933

For the above Bayesian network we obtained the Log-Likelihood value of **-1305.880.**

Interesting Conditional Probabilities

From the above Bayesian Network we can arrive at some interesting conditional probabilities.

We can derive the joint probability by:

P(CS_Score, ResearchOverhead, AdminBasePay, Tuition) =
P(CS_Score)*P(ResearchOverhead|CS_Score)
***P(AdminBasePay|Tuition)*P(Tuition|CS_Score)**

CS Score is the only independent variable here, others are all conditioned.

- From "University Data" and the scatter matrix shown above we can observe that CS Score influences Research overhead to some extent. Since it is positively correlated, one value increases when the other increase. But we do see some exception (University at Buffalo has a CS Score of 2.6 but has a research overhead of 59.5, which is third in the list). The values are not completely correlated but are correlated to some extent (0.456).
- There might be other factors which influence the variables, such as Tuition fees may depend on whether the university is private or public. So if we have more data and more variables we can obtain a more accurate Bayesian Network to answer our queries.

Implementation

Following is the code to implement conditional probability for a variable with 1 parent

```
def cond_prob1(list_y,list_x):
   x0=np.ones(49);
   x1=np.array(list_x)
   a=np.zeros((2,2))
   row=0;col=0;
   for i in x0,x1:
      for j in x0,x1:
         a[row][col]=(prod(j,i));
         col=col+1;
      row=row+1;
      col=0;
   y1=np.array(list_y);
   y00=prod(y1,x0);
   y10=prod(y1,x1);
   Y=[y00],[y10];
   Y=np.asmatrix(Y)
   A=np.asmatrix(a)
   A_inv = np.linalg.inv(A)
   beta=A_inv*Y
   sigma,X=sig1(beta,x1,y1)
   logterm=math.log(2*math.pi*sigma)
   log_pdf=(-24.5)*((logterm+1))
   return log_pdf
```

Following is the code to calculate sigma

```
def sig1(beta,x,y):
    sigma=0;
    z=[]
    for i in range(49):
        z.append(math.pow(((float(beta[0])+float(beta[1])*x[i])-y[i]),2))
        sigma+=z[i];
    sigma=sigma/49;
    return sigma,z;
```

Results

UBitName = sugoshna
personNumber = 50207357
mu1 = 3.214
mu2 = 53.386
mu3 = 469178.816
mu4 = 29711.959
var1 = 0.457
var2 = 12.850
var3 = 14189720820.903
var4 = 31367695.790
sigma1 = 0.676
sigma2 = 3.585
sigma3 = 119120.615
sigma4 = 5600.687
covarianceMat =
[[4.58000000e-01 1.10600000e+00 3.87978200e+03 1.05848000e+03]
[1.10600000e+00 1.28500000e+01 7.02793760e+04 2.80578900e+03]
[3.87978200e+03 7.02793760e+04 1.41897208e+10 -1.63685641e+08]
[1.05848000e+03 2.80578900e+03 -1.63685641e+08 3.13676958e+07]]
correlationMat =
[[1. 0.456 0.048 0.279]
[0.456 1. 0.165 0.14]
[0.048 0.165 1. -0.245]
[0.279 0.14 -0.245 1.]]
logLikelihood = -1315.119
BNgraph =
[[0 1 0 1]
[0 0 0 0]
[0 0 0 0]
[0 0 1 0]]
BNlogLikelihood = -1305.880