Simon Fraser University
Assignment 3 - CMPT 419/726: Machine Learning, Fall 2018

Date**: November 22nd, 2018**

Name: **Anurag Bejju**
Student ID: **301369375**
Professor: **Dr. Greg Mori**

---

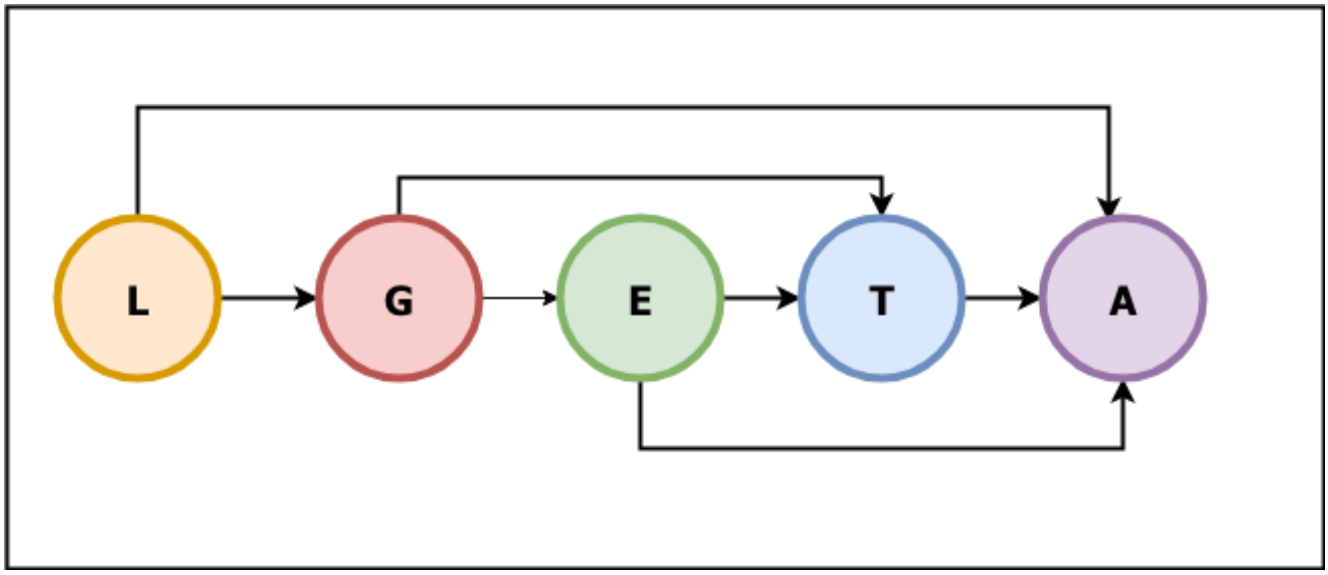
1. Graphical Models (22 marks)

1.1. Draw a simple Bayesian network for this domain.

Inference for creating the Bayesian network for this domain:

No.	Random Variable	Variable Type	Influence	Description
1	A	Boolean	-	$True$ - person will attend SFU $False$ - person will not attend SFU
2	L	Discrete	G	Parents' education level – Can have an influence on which party gets elected. • Parents having a university education (u) incline to vote for the liberal party (l) • Parents' with non-university education (o) incline to vote for NDP (d).
			A	Parents' education level might influence if the student attended university or not.
3	G	Discrete	E	The current provincial government will have an influence on the size of economy (eg: Taxes, regulation etc).
			T	The current provincial government will also have an impact on the price of tuition (eg: scholarships etc).
4	E	Continuous	T	Size of the economy will have an impact on the price of tuition.
			A	Size of the economy could also influence students decision to study in that province or not. Here $GDP\ (billions\ of\ \$)$ is used as a measuring factor.
5	T	Continuous	A	Cost of Tuition $(\$)$ is definitely be one of the major factors that could decide if the student attends SFU or not.

Bayesian Network:



1.2. The factored representation for the joint distribution $p(\mathrm{A}, \mathrm{L}, \mathrm{G}, \mathrm{E}, \mathrm{T})$ described by my Bayesian network is

$$\boldsymbol{p}(\mathbf{A}, \mathbf{L}, \mathbf{G}, \mathbf{E}, \mathbf{T}) = \boldsymbol{p}(\boldsymbol{L}) * \boldsymbol{p}(\boldsymbol{G} \mid \boldsymbol{L}) * \boldsymbol{p}(\boldsymbol{E} \mid \boldsymbol{G}) * \boldsymbol{p}(\boldsymbol{T} \mid \boldsymbol{G}, \boldsymbol{E}) * \boldsymbol{p}(\boldsymbol{A} \mid \boldsymbol{L}, \boldsymbol{E}, \boldsymbol{T})$$

1.3. Here are all the necessary conditional distributions for the above Bayesian network. I have also provided the type of distribution that should be used and have given a rough guidance / example values for parameters using educated guesses.

$\boldsymbol{p}(\boldsymbol{L})$

Input Type - Nil
Output Type – Discrete

Probability Distribution: According to 2016 Stats Canada Report [1], the Educational attainment of working-age population who got a university degree in 2016 is *24.7%.* So based on that, the distribution would be

$\boldsymbol{p}(\boldsymbol{L} = \boldsymbol{u})$	0.247
$\boldsymbol{p}(\boldsymbol{L} = \boldsymbol{o})$	0.753

$\boldsymbol{p(G \mid L)}$

Input Type: Discrete $(\boldsymbol{L})$
Output Type: Discrete $(\boldsymbol{G})$
Probability Distribution: According to ipolitics poll, Parents having a university education (u) incline to vote for the liberal party (l) *36%* of the time and Parents' with non-university education (o) incline to vote for NDP (d) *33%* of the time. So based on that, the distribution would be

	$\boldsymbol{L = o}$	$\boldsymbol{L = u}$
$\boldsymbol{G = Liberal\ (l)}$	0.19	0.36
$\boldsymbol{G = NDP\ (d)}$	0.33	0.12

$\boldsymbol{p(E \mid G)}$

Input Type: Discrete $(\boldsymbol{G})$
Output Type: Continuous $(\boldsymbol{E})$
Probability Distribution: This is a case of a continuous output with discrete parents. This would result to two different Gaussian distributions based on the parent's value. Based on Statista report, the GDP of BC when NDP formed government in 2017 was *218.76* billion ($) and the GDP of BC when Liberal formed government in 2016 was *219.553* billion ($) [3]. So based on that, the two possible normal distributions for size of the economy if given the current government.

$\boldsymbol{G = Liberal\ (l)}$	$\boldsymbol{G = NDP\ (d)}$
$\boldsymbol{p(E \mid G = l) = \mathcal{N}(\mu = 218.76, \sigma = 33\ )}$	$\boldsymbol{p(E \mid G = l) = \mathcal{N}(\mu = 219.55, \sigma = 26\ )}$

$\boldsymbol{p(T \mid E, G)}$

Input Type: Continuous $(\boldsymbol{E}\ )$ **,** Discrete $(\boldsymbol{G})$
Output Type: Continuous $(\boldsymbol{T})$
Probability Distribution: This is a case of a continuous output with a continuous and a discrete parent. This would result to two Linear Gaussian models for each current government type. So based on that, the distribution would be.

$\boldsymbol{G = Liberal\ (l)}$	$\boldsymbol{G = NDP\ (d)}$
$\boldsymbol{p(T \mid E, G) = \mathcal{N}(T; \mu = 276E + 20, \sigma = 45\ )}$	$\boldsymbol{p(T \mid E, G) = \mathcal{N}(T;\ \mu = 316E + 36, \sigma = 35\ )}$

$\boldsymbol{p(A \mid T, E, L)}$

Input Type: Continuous ($\boldsymbol{T}$) **,** Continuous ($\boldsymbol{E}$) **,** Discrete ($\boldsymbol{L}$)
Output Type: Binary ($\boldsymbol{A}$)

Probability Distribution: This is a case of a Binary output with two continuous and a discrete parent. This would result to two multi-variate sigmoid models for each discrete value of ($\boldsymbol{L}$). So based on that, the distribution for student attending SFU would be.

Distribution for student attending SFU (A = True)	
$\boldsymbol{L = o}$	$\boldsymbol{L = u}$
$\boldsymbol{p(A = True \mid T, E, L) = Sigmoid\ (100 + 840t + 300e)}$	$\boldsymbol{p(A = True \mid T, E, L) = Sigmoid\ (65 + 942t + 240e)}$

Where $t = tuition\ value\ (\$)\ and\ e = size\ of\ the\ economy = GDP\ (billions\ of\ \$)$

The distribution for student not attending SFU would be:

Distribution for student not attending SFU (A = False)	
$\boldsymbol{L = o}$	$\boldsymbol{L = u}$
$\boldsymbol{p(A = False \mid T, E, L) = 1 - Sigmoid\ (100 + 840t + 300e)}$	$\boldsymbol{p(A = False \mid T, E, L) = 1 - Sigmoid\ (65 + 942t + 240e)}$

Where $t = tuition\ value\ (\$)\ and\ e = size\ of\ the\ economy = GDP\ (billions\ of\ \$)$

**<u>Note:</u> In a scenario where we have Discrete Child with a Continuous Parent, we can represent it using a multi-variate sigmoid model. That is*

$$P(X = True|Y) = \frac{1}{1 + \exp(\boldsymbol{w}^T * \boldsymbol{y})}$$

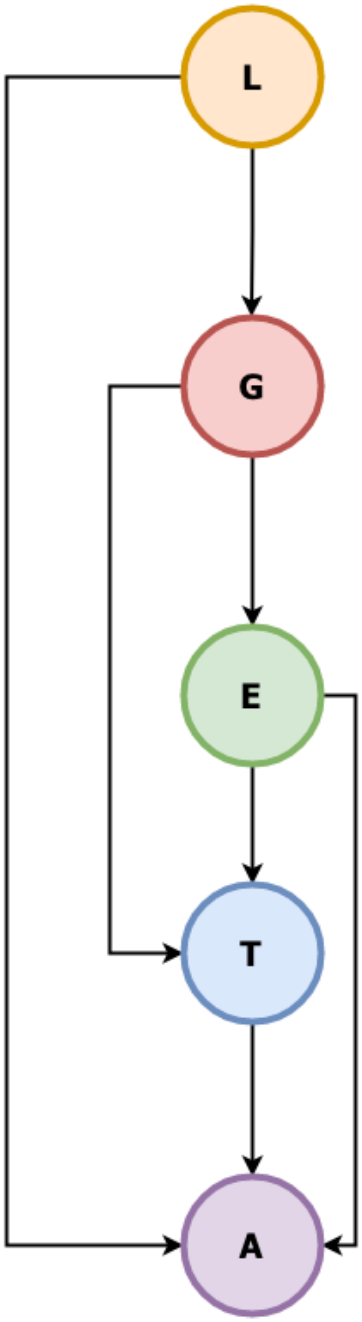
Therefore in the above case

$$P(A = True|T, E, L) = \frac{1}{1 + \exp(\boldsymbol{w}^T * \boldsymbol{x})}$$

where

$$w = \begin{pmatrix} 100 \\ 840 \\ 300 \end{pmatrix} and\ x = \begin{pmatrix} 1 \\ t \\ e \end{pmatrix}$$

Bayesian Network with Conditional Distributions



P (L)	
p (L = u)	0.247
p (L = o)	0.753

P (G | L)		
	L = o	L = u
G = Liberal	0.19	0.36
G = NDP	0.33	0.12

P (E | G)	
G = Liberal	G = NDP
p (E | G = l) = N(μ= 218.76, σ = 33)	p (E|G = o) = N(μ= 219.55, σ = 26)

P (T | E, G)	
G = Liberal	G = NDP
p (T | G,E) = N(T; 276E + 20, 45)	p (T|G,E) = N(T; 316E + 36, 35)

P (A | T,E,L) when A is TRUE	
L = u	L = o
p (A | T,E,L) = σ (100 + 840t + 300e)	p (A |T,E,L) = σ (65 + 942t + 240e)

Anurag Bejju (301369375)

1.4. As we know, *Maximum Likelihood Estimation (MLE)* is a method of estimating the parameters of a model, given observations. It attempts to find the parameter values that maximize the likelihood function, given the observations.
For a Bayesian Network, the Likelihood can be represented as

$$\boldsymbol{L(\theta : D) = \prod_{m=1}^{N} P(x_1[m] \ldots . x_k[m] : \theta) = \prod_i \prod_{m=1}^{N} P(x_i[m] \mid pa_i[m] : \theta) = \prod_i L_i(\theta_i : D)}$$

As the probability of each data point can be factored out based on the Bayesian network we created,

$$\begin{aligned} \boldsymbol{P(x_k)} &= \boldsymbol{p(\mathrm{A} = a_k, \mathrm{L} = l_k, \mathrm{G} = g_k, \mathrm{E} = e_k, \mathrm{T} = t_k)} \\ &= \boldsymbol{p(\mathrm{L} = l_k) * p(\mathrm{G} = g_k \mid \mathrm{L} = l_k) * p(\mathrm{E} = e_k \mid \mathrm{G} = g_k)} \\ &* \boldsymbol{p(\mathrm{T} = t_k \mid \mathrm{G} = g_k, \mathrm{E} = e_k) * p(A \mid \mathrm{L} = l_k, \mathrm{E} = e_k, \mathrm{T} = t_k)} \end{aligned}$$

Using this in our likelihood function, we get

$$\begin{aligned} \boldsymbol{L(\theta : D)} = \prod_{m=1}^{N} & \boldsymbol{p(\mathrm{L[m]} : \theta_L) * p(\mathrm{G[m]} \mid \mathrm{L[m]} : \theta_{G|L})} \\ & \boldsymbol{* p(\mathrm{E[m]} \mid \mathrm{G[m]} : \theta_{E|G}) * p(\mathrm{T[m]} \mid \mathrm{G[m]}, \mathrm{E[m]} : \theta_{T|G,E})} \\ & \boldsymbol{* p(\mathrm{A[m]} \mid \mathrm{L[m]}, \mathrm{E[m]}, \mathrm{T[m]} : \theta_{A|L,E,T})} \end{aligned}$$

Where m indicates the function goes over all the x values from $\boldsymbol{x_1 to\ x_N}$.
Let's consider the local likelihood function for A, then our :

$$\boldsymbol{L(x_a \mid pa(x_a)) = \prod_{m=1}^{N} p(\mathrm{A[m]} \mid \mathrm{L[m]}, \mathrm{E[m]}, \mathrm{T[m]} : \theta_{A|L,E,T})}$$

As we can infer from the above equation, the local likelihood functions arise do not depend on all the given parameters of $\boldsymbol{x}$. Therefore, we only maximize the likelihood of each parameter locally. When examining $\boldsymbol{P(A \mid parents(A))}$ we see that it is only a function of $A, L, E\ and\ T$. As a result, only $A_n, L_n, E_n\ and\ T_n$ parameters of x will be used. To learn the parameters $\theta_{A|L,E,T}$ we would have to maximize the likelihood function with respect to $\theta_{A|L,E,T}$. We can do this by taking the logarithm of the function and maximizing it.

2. KL Divergence (17 marks)

2.1. To Prove: The difference between a probability distribution and itself is 0.
(i.e $D_{KL}(P||P) = 0$

Given: The Kullback-Leibler (KL) divergence is a measure of the difference from one probability distribution P to another Q. It is denoted by and is defined as:

$$D_{KL}(P||Q) = \int P(x) ln \frac{P(x)}{Q(x)} dx$$

for continuous random variable X

Based on the above equation we get:

$$D_{KL}(P||P) = \int P(x) \ln(\frac{P(x)}{P(x)}) dx$$

Since Jensen Inequality states:

$$\int \ln\big(f(x)\big)\, dx \geq \ln(\int f(x))$$

We can transform the above equation to

$$D_{KL}(P||P) = \ln\left(\int P(x) \frac{P(x)}{P(x)} dx\right)$$

$$= \ln\left(\int P(x)\, dx\right)$$

Since the integral of a probabilistic distribution is one, we get

$$\boxed{D_{KL}(P||P) = \ln(1) = 0}$$

2.2. No KL divergence is not symmetric.

i.e $D_{KL}(P||Q) = D_{KL}(Q||P)$ *is not necessarily true*

2.3. To Prove: Suppose $\boldsymbol{\mu}_p = \boldsymbol{\mu}_q$, Which KL divergence is larger, $D_{KL}(P||Q)$ or $D_{KL}(Q||P)$.

Given: The (KL) divergence between a pair of Gaussian distributions:

$$p(x) = \mathcal{N}\left(x\,;\, \mu_p\, \sigma_p^2\right) \; and \; q(x) = \mathcal{N}\left(x\,;\, \mu_q\, \sigma_q^2\right)$$

The formula for the KL divergence between two univariate Gaussian distributions is

$$D_{KL}(P||Q) = \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{\sigma_p^2 + \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right)^2}{2\sigma_q^2} - \frac{1}{2}$$

Since $\boldsymbol{\mu}_p = \boldsymbol{\mu}_q$:

$$D_{KL}(P||Q) = \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2}$$

$$D_{KL}(P||Q) + \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{1}{2} = 2 * \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{\sigma_p^2}{2\sigma_q^2}$$

Similarly:

$$D_{KL}(Q||P) = \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2}$$

$$D_{KL}(Q||P) - \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{1}{2} = \frac{\sigma_q^2}{2\sigma_p^2}$$

Since :

$$\frac{x}{2} > \ln(x) + \frac{1}{2x} \text{ when } x > 1$$

if $x = \frac{\sigma_q^2}{\sigma_p^2}$ in the above equation then $\frac{\sigma_q^2}{\sigma_p^2} > 1 \Rightarrow \sigma_q^2 > \sigma_p^2$ and:

$$\frac{\sigma_q^2}{2\sigma_p^2} > \ln\left(\frac{\sigma_q^2}{\sigma_p^2}\right) + \frac{\sigma_p^2}{2\sigma_q^2} \Rightarrow \frac{\sigma_q^2}{2\sigma_p^2} > 2 * \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q^2}{2\sigma_p^2}$$

substituting equation 1 and 2 we get

$$D_{KL}(Q||P) - \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{1}{2} > D_{KL}(P||Q) + \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{1}{2}$$

$$D_{KL}(Q||P) - \ln(\sigma_p) + \ln(\sigma_q) > D_{KL}(P||Q) + \ln(\sigma_q) - \ln(\sigma_p)$$

Similary

$\boldsymbol{D_{KL}(Q\|\|P) > D_{KL}(P\|\|Q)}$	$\boldsymbol{if\ \sigma_q^2 > \sigma_p^2}$
$\boldsymbol{D_{KL}(P\|\|Q) > D_{KL}(Q\|\|P)}$	$\boldsymbol{if\ \sigma_p^2 > \sigma_q^2}$

3. Gated Recurrent Unit (10 marks)

3.1. What values of r_j and z_j would cause the new state for h_j to be similar to its old state? Give a short, qualitative answer.

z_j is the UPDATE gate. It decides how much information from the past must be let through.
r_j is the RESET gate. It decides how much information from the past must be discarded.
h_j is Final memory at time step combines current and previous time steps.
$\tilde{h}_j$ is New memory content / new hidden state.

The equation for the new state h_j is

$$h_j = z_j h_j^{(t-1)} + (1 - z_j)\tilde{h}_j^t$$

Based on the above equation, if the value of z_j is close to one then $(1 - z_j)\tilde{h}_j^{(t-1)}$ gets close to zero and h_j will be similar to its old state $h_j^{(t-1)}$. Since the value of r_j doesn't affect the outcome of h_j when z_j is close to 1, it can take any value between 0 and 1 [As the value is squashed by sigmoid function].

3.2. If r_j and z_j are both close to 0, how would the state for h_j be updated? Give a short, qualitative answer.
When r_j and z_j are both close to 0, the new state h_j tends to take the value present in $\tilde{h}_j^t$ (new hidden state).
Since

$$\tilde{h}_j^t = \emptyset(W_j x_j + [U_j\ (r_j \circ h_j^{(t-1)})])$$

and r_j is close to zero, then $\tilde{h}_j^t$ gets the value of $\emptyset(W_j x_j)$ where x_j is the input vector. Therefore the state of h_j will be updated by the input vector x_j and the past state information would be discarded.

References:

[1] https://www150.statcan.gc.ca/n1/pub/12-581-x/2017000/edu-eng.htm

[2] https://ipolitics.ca/2015/05/29/the-ekos-poll-canadians-warming-to-the-idea-of-coalition-government/

[3] https://www.statista.com/statistics/577563/gdp-of-british-columbia-canada/