1 Graphical Models

1.1 Draw a simple Bayesian network for this domain

A Bayesian network has been drawn based on the below justification: -

Taken, a boolean random variable

A - person attending SFU - t/f.

L - max of parents education level - o/u.

G - current provincial govt - l/d.

E - current provincial economy size

T - SFU tuition level.

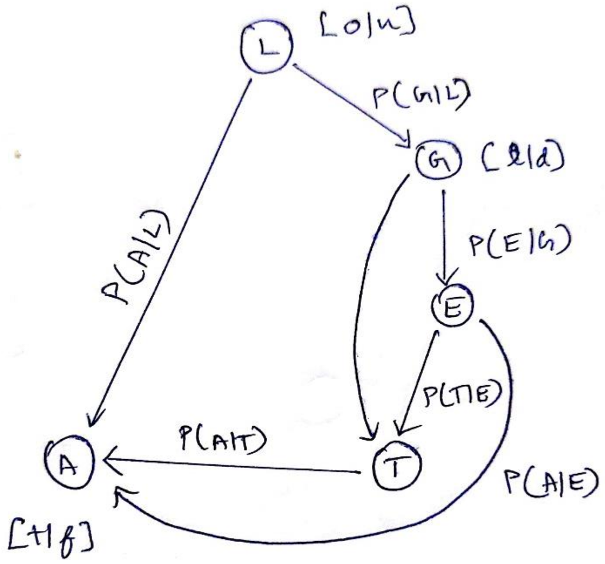
Given the above, we get the following conditional dependencies.

$IL \rightarrow G$: The parent's education level likely has an influence on who they voted for. Parents with university education likely voted for the liberal party due to their middle wing ideologies. Parents with no university education likely voted for NDP due to their left wing ideologies.

$G \rightarrow E$: The government has an influence on the size of economy. We should have some sort of dependency. But, it is most likely independent of whether parents went to university or not, given the current government is in power.

G → T
E → T : The current government and the size of the economy will have an impact on the price of tuition.

L → A
E → A
T → A : Whether or not the student attends SFU depends on the cost of tuition, and the parents will also influence the student depending on whether or not they attended university. The size of the economy might relate to how many students are admitted to SFU.



1.2) Using the Bayesian network constructed above, the factored joint distribution is:

Factored representation of the joint distribution $P(A, L, G, E, T)$ that is described in my Bayesian network.

$$P(A, L, G, E, T) = P(G) \cdot P(L) \cdot P(E|G) \cdot P(T|E, G) \cdot P(A|L, T, E)$$

Have derived this using the product rule of probability.

According to the product rule, it states

$$P(x_1, x_2, \ldots x_n) = P(x_n | x_1, x_2 \ldots x_{n-1}) \ldots P(x_2|x_1) P(x_1)$$

Thus applying the product rule in my statement, we get

$$P(G) \cdot P(L) \cdot P(E|G) \cdot P(T|E, G) \cdot P(A | L, T, E)$$

1.3) Considering the parents are either discrete or continuous, the below probability distributions were derived:

a. P(L) = Discrete output. Using an educated guess:

P(L = u)	0.4
P(L = o)	0.6

b. P(G|L) = Discrete outputs. Using an educated guess:

	L = o	L = u
G = Liberal	0.3	0.1
G = NDP	0.2	0.4

c. P(E|G) = continuous output with discrete parents. The resulting system could have 2 different Gaussian distributions based on the values of the parents. The table below gives two possible normal distributions of the GDP (measured in billions) of British Columbia given the current government.

G = liberal	G = NDP
P(E|G = liberal) = N(μ = 275, σ = 20)	P(E|G = NDP) = N(μ = 250, σ = 15)

d. P(T|E,G) = continuous output with discrete parents and continuous parents. The resultant distribution could include 2 linear Gaussians. One for G = liberal and one for G = NDP. As the size of the economy increases, the tuition might also increase and there is a probability associated with this point. One possibility is shown below in the table.

G = liberal	G = NDP
P(T|E, G) =N(T; 3000E; 100)	P(T|E, G) =N(T; 4000E; 120)

e. P(A|L, A, T) = Discrete output with continuous parents and discrete parents. To model this distribution we can use a two multi-variate sigmoids. One for L = u and and for L = o. This would allow for two continuous inputs and the value of the sigmoid could be used to give a discrete output. One possibility is shown below. The following equations assume:

i. If the parents went to university, the students are more likely to go to university regardless of the price and the current state of the economy.

ii. An increase in tuition will decrease the probability of the student attending university.
iii. If the economy is doing well, people are more likely to invest in an education.

For L = u:

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

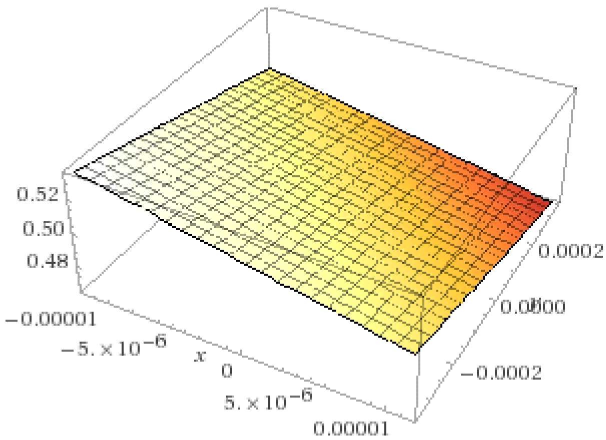
Where:

$$\boldsymbol{w} = \begin{pmatrix} 6000 \\ 250 \end{pmatrix}$$

$$\boldsymbol{x} = \begin{pmatrix} t \\ e \end{pmatrix}$$

- $t = tuition\ (\$)$
- $e = GDP\ (billions\ of\ \$)$

The resulting equation was plotted using Wolfram Alpha:



For L = o:

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

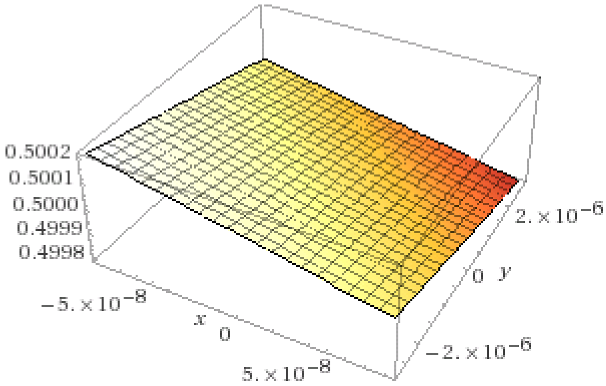
Where:

$$\boldsymbol{w} = \begin{pmatrix} 5000 \\ 230 \end{pmatrix}$$

$$\boldsymbol{x} = \begin{pmatrix} t \\ e \end{pmatrix}$$

- $t = tuition$ ($)
- $e = GDP$ (billions of $)

The equation was then plotted using Wolfram Alpha:



1.4 To find the maximum likelihood estimates for the parameters, the maximum likelihood equation must first be generated. This function is given below. The maximum likelihood estimate examines the probability of the data, given the observed training set.

$$L(\theta : D) = P(D|\theta) = \prod_{k=1}^{N} P(x_k : \theta)$$

The probability of each data point can be factored out using the Bayesian network above to give:

$$P(x_k) = P(A = a_k, L = l_k, E = e_k, T = t_k, G = g_k)$$

$$P(x_k) = P(L = l_k) P(G = g_k | L = l_k) P(E = e_k | G = g_k) \ldots$$

$$\ldots P(T = t_k | E = e_k, G = g_k) P(A = a_k | L = l_k, E = e_k, T = t_k)$$

Plugging this back into the maximum likelihood equation we obtain:

LOCAL LIKELIHOOD FUNCTIONS

$$L(D:\theta) = \prod_{m} \overbrace{P(A[m] : \theta_A)}\, \overbrace{P(G[m] \mid L[m] : \theta_{G|M})} \cdots$$

$$\cdots P(E[m] \mid G[m] = \theta_{E|G})\, P(T[m] \mid E[m], G[m] = \theta_{T|E,G}) \cdots$$

$$\cdots P(A[m] \mid L[m], T[m], E[m] - \theta_{A|L,T,E})$$

Where m indicates the function spans over all the x values from $x_1 to\ x_m$. The local likelihood function for A is then:

$$L(x_A \mid Pa(x_A)) = \prod_{M} P(A[M] \mid L[M], T[M], E[M] = \theta_{A|L,T,E})$$

When we plug these equations back into the likelihood equation, we see that local likelihood functions arise which do not depend on all the given parameters of **x**.

- As a result, we only have to maximize the likelihood of each parameter locally. When examining P(A|parents(A)) we see that it is only a function of $a, l, t, and\ m$. As a result, only the a_n, t_n, e_n, l_n parameters of **x** are needed.
- To learn the parameters for $\theta_{A|L,T,E}$ one would have to maximize the likelihood function with respect to $\theta_{A|L,T,E}$.
- This would likely involve taking the logarithm of the function and maximizing that by taking the gradient and solving where it equals 0.

2 KL Divergence

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

- Show $D_{KL}(P \| P) = 0$, applying to above, we derive

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

$$= \int P(x) \ln(1) dx = \int P(x)\, 0\, dx$$

Since $\ln(1) = 0 \Rightarrow 0$

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

(2.) No. KL divergence is not symmetric.

(3.) Gaussian Distribution 1 Gaussian Distribution 2

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

Formula

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

- $\mu_q = \mu_p$ $\rightarrow$ (1)
- $x/2 > \ln(x) + 1/2x$ for $x > 1$ $\rightarrow$ (2).

While we start to derive,

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

(By using (1))

$$= \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2} \quad — (I)$$

Also, when we derive $D_{KL}(Q||P)$, we get

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

$$= \ln \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2} \quad — (II)$$

From (I) & (II), we now find which is larger. $D_{KL}(P||Q)$ (or) $D_{KL}(Q||P)$

Subtract (I) - (II)

$$\Rightarrow D_{KL}(P||Q) - D_{KL}(Q||P)$$

$$= \left[\ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2}\right] - \left[\ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2}\right]$$

$$= \left[\ln\left(\frac{\sigma_q}{\sigma_p}\right) - \ln\left(\frac{\sigma_p}{\sigma_q}\right)\right] + \left[\frac{\sigma_p^2}{2\sigma_q^2} - \frac{\sigma_q^2}{2\sigma_p^2}\right] \cancel{- \frac{1}{2}} \cancel{+ \frac{1}{2}}$$

Let $\frac{\sigma_q^2}{\sigma_p^2} = x$, as given we can derive the above as,

$$= \ln \frac{\sigma^2 q}{\sigma^2 p} + \frac{\sigma^2 p}{2\sigma^2 q} - \frac{\sigma^2 q}{2\sigma^2 p}$$

$$= \ln(x) + \tfrac{1}{2}\left[\tfrac{1}{x} - x \right]$$

Since it is given that $x/2 > \ln(x) + \frac{1}{2x}$ for $x > 1$

We can change the following to

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

→ positive outcome for $x > 1$

Therefore $D_{KL}(Q \| P)$ is larger than $D_{KL}(P \| Q)$.

3 Gated Recurrent Unit

(3) Find the values of r_j and z_j making the new state of h_j similar to its old state.

If $z_j = 1$, Then $h_j^{(t)}$ will be equal to its old state. The GRU equations can be written as :-

$$r_j = \sigma\left([W_r x]_j + [U_r h(t-1)]_j\right) \rightarrow (1).$$

$$z_j = \sigma\left([W_z x]_j + [U_z h(t-1)]_j\right) \rightarrow (2).$$

$$h_j^{(t)} = z_j h_j^{(t-1)} + (1-z_j)\tilde{h}_j^{(t)} \rightarrow (3)$$

$$\tilde{h}_j^{(t)} = \phi\left([Wx]_j + [U(r \odot h(t-1))]\right] \rightarrow (4)$$

With the above equations, we can now deduce that if $z_j = 1$

From eq (3), $\boxed{h_j^{(t)} = \tilde{h}_j(t)}$.

(ii) If r_j & z_j are close to 0, how would the state for h_j be updated.

For z_j close to 1, $h_j^{(t)}$ will not be the same as old state, but it will be similar to its old state.

r_j can be any value, as its value effects $\tilde{h}_j^{(t)}$ which would be multiplied by $z_j = 1$

Thus, for $z_j = 1$ & r_j = any value that can be between (0 to 1) for the new state, h_j to be similar to its old state.

(b)

$$r_j = \sigma([W_r x]_j + [U_r h(t-1)]_j) \rightarrow (1)$$

$$z_j = \sigma([W_z x]_j + [U_z h(t-1)]_j) \rightarrow (2)$$

$$h_j^{(t)} = z_j h_j^{(t-1)} + (1 - z_j)\tilde{h}_j^{(t)} \rightarrow (3)$$

$$\tilde{h}_j^{(t)} = \phi([Wx]_j + [U(r \odot h(t-1))]) \rightarrow (4)$$

W_r & U_r are weights matrices which are learned.

When r & z are both close to 0, the hidden state is forced to ignore the previous hidden state & reset with the current input. As z is close to 0, it will prevent the previous hidden state to be updated by the new hidden state.

$$r_j = 0 \; ; \; z_j = 0$$

Therefore eq (3) becomes $\boxed{\tilde{h}_j^{(t)} = \phi([Wx]_j)}$