

CMPT-726

Assignment 3: Graphical Models / Recurrent Neural Networks

Due November 15 at 11:59pm

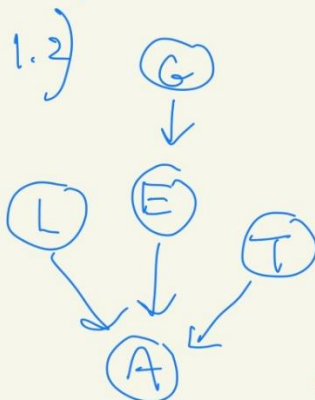
1. Graphical Models (22 marks)

- 1.1. Draw a simple Bayesian network for this domain.
- 1.2. Write the factored representation for the joint distribution $p(A, L, G, E, T)$ that is described by your Bayesian network.

ANS:

①: Det. whether local high school student will attend SFU
 var. 1: $A \rightarrow \text{true/false}$: attend SFU
 2: $L \rightarrow o, u$: max. parents' education level
 discrete } (non-...) (university)
 var. } 3: $G \rightarrow l, d$: current provincial govt.
 (liberal) (mop)
 cont. } 4: $E \rightarrow \text{cont. num}$: current provincial econ size
 var. } 5: $T \rightarrow \text{cont. num}$: SFU tuition level

1.1)
$$p(A, L, G, E, T) = P(G)P(E|G)P(L)P(T)P(A|L, E, T)$$



⇒ Logic behind these :

- $P(A|L)$: parents' education level could help determine whether their children will follow their parent or not (as parent with university level is likely to convince their children to do the same)
- $P(A|E)$: good provincial economy size means good job market which in turn allows local people to get good jobs after their graduation; thus, local high school students are likely to go to their local university (SFU).
- $P(E|G)$: good current provincial government is responsible for good current provincial economy size.

1.3. Supply all necessary conditional distributions. Provide the type of distribution that should be used and give rough guidance / example values for parameters (do this by hand, educated guesses).

ANS:

1.3) $\Rightarrow P(E|G) \& P(T)$: continuous rv.

I use Gaussian distribution with mean (μ) and variance (σ^2) as parameter because it has common occurrence in many natural phenomena (so it is common and safe to assume the unknown distributions to be Gaussian distributions).

$\Rightarrow P(L), P(G), P(A|L,E,T)$: discrete random var.

I use Bernoulli distribution.

parameter p is probability of event 1 occurs while $1-p$ is a probability of event 2 occurs.

For example, $P(G)$ with $p = 0.8$ is probability that current provincial government is MPD w/ 80% prob while it is Liberal w/ prob of 20%.

1.4. Suppose we had a training set and wanted to learn the parameters of the distributions using maximum likelihood. Denote each of the N examples with its values for each random variable by $x_n = (a_n, l_n, g_n, e_n, t_n)$. The training set is $\{x_1, x_2, \dots, x_N\}$. Which elements of the training data are needed to learn the parameters for $p(A|paA)$? Why? Start by writing down the likelihood and argue from there.

ANS:

1.4)

parameters: (Θ)

$\hookrightarrow \theta_G, \theta_L, \theta_T, \theta_{E|G}, \theta_{A|L,E,T}$

\Rightarrow Likelihood (L)

$L(\Theta; D) =$

$= \prod_{n=1}^N P(a_n, l_n, g_n, e_n, t_n; \Theta)$

$= \left(\prod_{n=1}^N P(a_n | l_n, e_n, t_n; \theta_{A|L,E,T}) \right) \times$

$\left(\prod_{n=1}^N P(l_n; \theta_L) \right) \left(\prod_{n=1}^N P(t_n; \theta_T) \right) \times$

$\left(\prod_{n=1}^N P(e_n | g_n; \theta_{E|G}) \right) \times$

$\left(\prod_{n=1}^N P(g_n; \theta_G) \right)$

Likelihood

In order to find the likelihood of this Bayesian Network, since the sets of parameters in conditional prob distribution are disjoint, total likelihood can be computed by a product of local likelihood functions, one per variable.

2. KL Divergence (20 marks)

2.1. ANS:

2.2. ANS:

② KL Divergence

2.1) Ans: KL divergence is not symmetric

$D_{KL}(P||Q) - D_{KL}(Q||P)$ may not be zero.

Even though the KL divergence measures the difference between 2 distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. ✕

2.2) Ans Show $D_{KL}(P||P) = 0$

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

$$= \int P(x) [\ln P(x) - \ln P(x)] dx$$

$$= \int P(x) \ln P(x) dx - \int P(x) \ln P(x) dx$$

$$= 0 \quad \text{✕}$$

2.3. ANS:

2.3) K_L is always non-negative.

Ans prove $D_{KL}(P||Q) \geq 0$ or $-D_{KL}(P||Q) \leq 0$

$$\begin{aligned} -D_{KL}(P||Q) &= -\int P(x) \ln \frac{P(x)}{Q(x)} dx \\ &= \int P(x) \ln \frac{Q(x)}{P(x)} dx \end{aligned}$$

from Jensen's inequality: \log/\ln is a concave function

$$\begin{aligned} -D_{KL}(P||Q) &= \int P(x) \ln \frac{Q(x)}{P(x)} dx \leq \ln \int P(x) \frac{Q(x)}{P(x)} dx \\ &\leq \ln 1 \end{aligned}$$

$$-D_{KL}(P||Q) \leq 0$$

$$D_{KL}(P||Q) \geq 0 \quad \text{or} \quad \text{#}$$

2.4. ANS:

$$\begin{aligned}
 2.4) \rightarrow D_{KL}(P||Q) &= \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \\
 &= \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2} = \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 - \sigma_q^2}{2\sigma_q^2} \\
 \rightarrow D_{KL}(Q||P) &= \ln \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2} = \ln \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 - \sigma_p^2}{2\sigma_p^2}
 \end{aligned}$$

note: $\frac{x}{2} > \ln(x) + \frac{1}{2x}$ for $x > 1$

$$\frac{\sigma_q}{2\sigma_q} > \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q}{2\sigma_p} \quad \left(\text{let } x = \frac{\sigma_p}{\sigma_q}\right)$$

for $\sigma_p > \sigma_q$

$$0 > \ln \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 - \sigma_p^2}{2\sigma_q\sigma_p}$$

$$\begin{aligned}
 0 &> D_{KL}(Q||P) \rightarrow \\
 0 &< D_{KL}(P||Q)
 \end{aligned}$$

∴ if $\sigma_p > \sigma_q$, $D_{KL}(P||Q) > D_{KL}(Q||P)$

$\sigma_q > \sigma_p$, $D_{KL}(Q||P) > D_{KL}(P||Q)$



3. Gated Recurrent Unit (10 marks)

3.1.ANS:

3.2.ANS:

$$\textcircled{3} \quad r_j = \sigma([W_r x]_j + [U_r h_{<t-1>}]_j)$$

$$z_j = \sigma([W_z x]_j + [U_z h_{<t-1>}]_j)$$

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

$$h_j^{<t>} = z_j h_j^{<t-1>} + (1 - z_j) \phi([W x]_j + [U(r \odot h_{<t-1>})]_j)$$

3.1) we want $h_j^{<t>} = h_j^{<t-1>}$

$\therefore z_j$ must be 1 or close to 1.
while if z_j is not equal to 1,
 r_j must be some number that
is close to 1 ($r \odot h_{<t-1>}$ as to
preserve $h_{<t-1>}$) ~~###~~

3.2) if r_j, z_j are zero.

$$h_j^{<t>} = 0 + (1) \phi([W x]_j)$$

$$h_j^{<t>} = \phi(\underbrace{[W x]_j}_{\text{current content}})$$

→ the hidden will use just current content.
(incoming new input x).

note: $z \Rightarrow$ update gate : helps model
determine how much the past information
(from previous time steps) is needed to be passed
to the future.

\therefore high $z \rightarrow$ preserve much past info
(h_{t-1})

note: $r \Rightarrow$ reset gate : helps model
determine how much the past information
to forget.

\therefore high $r \rightarrow \odot$ product will preserve
($r \odot h_{t-1}$) past information.

4. Attention Models (10 marks)

4.1.ANS:

④ 4.1) The purpose of the sinusoidal encoding is because the Transformer model does not require word inputs that will be fed into network to have specific orders or positions. Therefore, it needs some technique (sinusoidal encoding) to help the model incorporate the order of words by adding a position-dependent signal to each word-embedding.

→ The sinusoidal encoding does not require the sentence to be fix length compared to one-hot encoding scheme. It uses properties of $\sin(x)$ and $\cos(x)$ [cyclic] functions to return information of the position of a word in a sentence.

4.2.ANS:

$$4.2) \quad PE(pos, z_i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, z_i + 1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Formula for positional encoding

→ same position

$$\frac{PE(pos 1)}{PE(pos 1)} = \frac{PE(pos 2)}{PE(pos 2)}$$

$$\frac{\sin(pos 1)}{\cos(pos 1)} = \frac{\sin(pos 2)}{\cos(pos 2)}$$

$$\tan(pos 1) = \tan(pos 2) \quad \text{X}$$