Assignment 3: Graphical Models / Recurrent Neural Networks

Due November 15 at 11:59pm

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment

---

1 Graphical Models (22 marks)

Consider the problem of determining whether a local high school student will attend SFU or not. Define a boolean random variable A (true if the person will attend SFU), discrete random variables L (maximum of parents' education level: can take values o for non-university or u for university) and G (current provincial government: l for Liberal Party, d for NDP), and continuous valued random variables E (current provincial economy size) and T (SFU tuition level).

1. **4 marks**. Draw a simple Bayesian network for this domain.
2. **2 marks**. Write the factored representation for the joint distribution $p(A, L, G, E, T)$ that is described by your Bayesian network.
3. **8 marks**. 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).
4. **8 marks**. 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 $\boldsymbol{x}_n = (a_n, l_n, g_n, e_n, t_n)$. The training set is $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$.

   Which elements of the training data are needed to learn the parameters for $p(A|pa_A)$? Why?[1]

   Start by writing down the likelihood and argue from there.

2 KL Divergence (20 marks)

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

$$D_{KL}(P||Q) = \int P(x) \ln \frac{P(x)}{Q(x)} dx \tag{1}$$

for continuous random variable X (see PRML for more details).

- **3 marks** Is the KL divergence symmetric? (Yes or no.)
- **3 marks** Show $D_{KL}(P||P) = 0$. I.e. the difference between a probability distribution and itself is 0.
- **7 marks** Use the fact that $\ln(1+x) \leq x$ to prove that KL divergence is always non negative.
- **7 marks** Consider the KL divergence between a pair of Gaussian distributions: $p(x) = \mathcal{N}(x; \mu_p, \sigma_p^2)$ and $q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2)$. The formula for the KL divergence between two univariate Gaussian distributions is

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

[1] pa_A = parents of A

Suppose $\mu_q = \mu_p$. Which KL divergence is larger, $D_{KL}(P||Q)$ or $D_{KL}(Q||P)$? Analyze, in terms of the respective $\sigma_\cdot$ values.

You may use the fact that $\frac{x}{2} > \ln(x) + \frac{1}{2x}$ for $x > 1$.

3 Gated Recurrent Unit (10 marks)

A Gated Recurrent Unit (GRU) is another type of recurrent neural network unit with the ability to remember and forget components of the state vector[2].

Read Sec. 2.3 of the linked paper for the description of the GRU. Note that the GRU's state consists of a vector of $\mathbf{h}$ values. There are two gates, r_j and z_j, which control the update of h_j, the j^{th} component of the GRU state.

- 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.
- 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.

4 Attention Models (10 marks)

As an alternative to recurrent neural network structures, attention models can be used to analyze an input sequence directly to compute a sequence of output state representations.

Read Sec. 3.5 of the paper by Vaswani et al. at NIPS 2017 `https://arxiv.org/abs/1706.03762`.

- What is the purpose of the sinusoidal positional encoding? What's its advantage over one-hot encoding(e.g. encoding first position as $[1, 0, \ldots]$ etc.)?
- When would two positions get the same encoding?(Assume pos are integers.)

[2]Cho et al. EMNLP 2014 `https://arxiv.org/abs/1406.1078`

Submitting Your Assignment

The assignment must be submitted online at `https://courses.cs.sfu.ca`. You must submit one file:

1. An assignment report in **PDF format**, called `report.pdf`. This report should contain the solutions to questions 1-4.