Assignment 2: Neural Networks, Graphical Models

Due Nov 6 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 search engines to find solutions for any part of the 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.

Submitting Your Assignment

The assignment must be submitted online at `https://canvas.sfu.ca/`. You must submit the following file:

1. An assignment report in **PDF format**, named `report.pdf`. This report should contain the solutions to questions 1-3.
2. A compressed file, named `code.zip`. This should contain a directory Q1 which has the code and the readme file.

---

1 Neural Networks (5 marks)

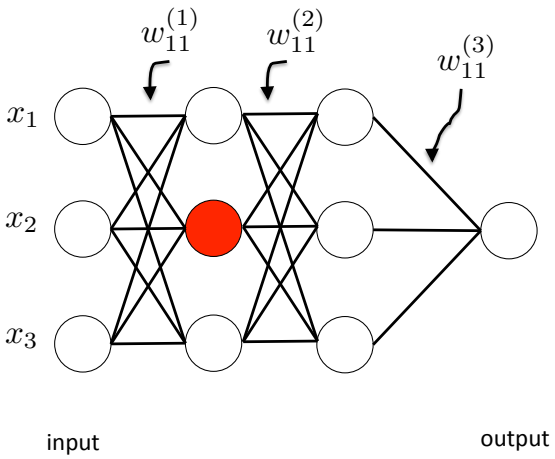
Error Propagation [3 marks]

We will derive error derivatives using back-propagation on the network below.

Notation: Please use notation following the examples of names for weights given in the figure. For activations/outputs, the red node would have activation $a_2^{(1)} = w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{23}^{(1)}x_3$ and output $z_2^{(1)} = h(a_2^{(1)})$.

Activation functions: Assume the activation functions $h(\cdot)$ for the hidden layers are logistics i.e $h(a) = 1/(1+e^{-a})$. For the final output node assume the activation function is an identity function $h(a) = a$.

Error function: Assume this network is doing regression, trained using the standard squared error so that $E_n(w) = \frac{1}{2}(y(\boldsymbol{x}_n, w) - t_n)^2$.



[1 mark] Consider the output layer.

- Calculate $\frac{\partial E_n(w)}{\partial a_1^{(3)}}$. Note that $a_1^{(3)}$ is the activation of the output node, and that $\frac{\partial E_n(w)}{\partial a_1^{(3)}} \equiv \delta_1^{(3)}$.
- Use this result to calculate $\frac{\partial E_n(w)}{\partial w_{12}^{(3)}}$.

[1 mark] Next, consider the penultimate layer of nodes.

- Write an expression for $\frac{\partial E_n(w)}{\partial a_1^{(2)}}$. Use $\delta_1^{(3)}$ in this expression.
- Use this result to calculate $\frac{\partial E_n(w)}{\partial w_{11}^{(2)}}$.

[1 mark] Finally, consider the weights connecting from the inputs.

- Write an expression for $\frac{\partial E_n(w)}{\partial a_1^{(1)}}$. Use the set of $\delta_k^{(2)}$ in this expression.
- Use this result to calculate $\frac{\partial E_n(w)}{\partial w_{11}^{(1)}}$.

Fine-Tuning a Pre-Trained Network [2 marks]

In this question you will experiment with fine-tuning a pre-trained network. This is a standard workflow in adapting existing deep networks to a new task.

We will utilize PyTorch (`https://pytorch.org`) a machine learning library for python.

The provided code builds upon ResNet 50, a state of the art deep network for image classification. ResNet 50 has been designed for ImageNet image classification with 1000 output classes.

The ResNet 50 model has been adapted to solve a (simpler) different task, classifying an image as one of 10 classes on CIFAR10 dataset.

The code `cifar_finetune.py` does the following:

- Constructs a deep network. This network starts with ResNet 50 up to its average pooling layer. Then, a small network with 32 hidden nodes then 10 output nodes (dense connections) is added on top.
- Initializes the weights of the ResNet 50 portion with the parameters from training on ImageNet.
- Performs training on only the new layers using CIFAR10 dataset – all other weights are fixed to their values learned on ImageNet.

The code can be downloaded from Canvas. You can either use Google Colaboratory (`https://colab.research.google.com/`) to run the code or setup virtual environment in your local. If you are setting up virtual environment in local, Anaconda (`https://www.anaconda.com`) environment config files with the latest stable release of PyTorch and torchvision are provided for Python 3.8 for Linux and macOs users. You can use the config file to create virtual environments and test your code. To set up the virtual environment, install Anaconda and run the following command

```
conda env create -f CONFIG_FILE.
```

Replace `CONFIG_FILE` with the path to the config file you downloaded. To activate the virtual environment, run the following command

```
source activate ENV_NAME
```

Replacing `ENV_NAME` with `cmpt726-pytorch-python38`.

Windows users please follow the instructions on PyTorch website (`https://pytorch.org`) to install manually. PyTorch only supports Python3 on Windows!

If you wish to download and install PyTorch by yourself, you will need PyTorch (v 0.4.1), torchvision (v 0.2.1), and their dependencies.

What to do:

Start by running the code provided. It will be *very* slow to train since the code runs on a CPU. You can try figuring out how to change the code to train on a GPU if you have a good GPU and want to accelerate training. Do the following tasks:

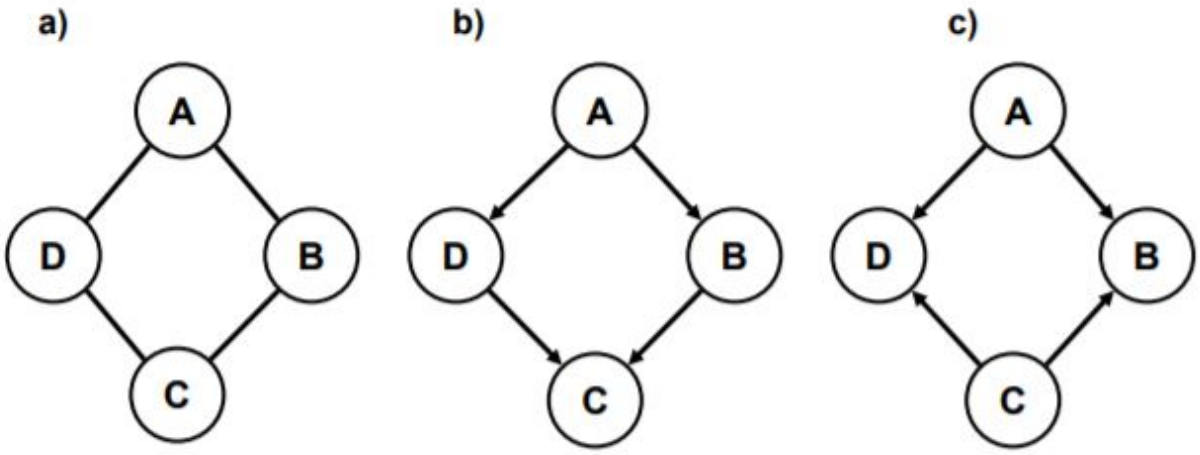
- Run validation of the model every few training epochs on validation or test set of the dataset and save the model with the best validation error. Report the best validation error and the corresponding training epoch in your report. (Do not submit saved models for this task.)
- Try applying $L2$ regularization to the coefficients in the small networks we added.

Put your code and a readme file for Question 1 under a separate directory named Q1 in the code.zip file you submit for this assignment. The readme file should describe what you implemented for the questions. It should also include the command to run your code.

2 Bayesian Networks (5 marks)

(i) [**1.5 marks**] Are the following statements true for each of the below networks. Please give your reasoning.

- $A \perp\!\!\!\perp C|B, D$ [**0.75 marks**]
- $B \perp\!\!\!\perp D|A, C$ [**0.75 marks**]



(ii) [**3.5 marks**] A patient goes to the doctor for a medical condition, the doctor suspects three diseases as the cause of the condition. The three diseases are D1, D2, D3, which are marginally independent from each other. There are four symptoms S1, S2, S3, S4 which the doctor wants to check for presence in order to find the most probable cause of the condition. The symptoms are conditionally dependent to the three diseases as follows: S1 depends only on D1, S2 depends on D1 and D2. S3 is depends on D1 and D3, whereas S4 depends only on D3. Assume all random variables are Boolean, they are either 'true' or 'false'.

a) [**1 mark**] Draw the Bayesian network for this problem.

b) [**0.5 mark**] Write down the expression for the joint probability distribution as a product of conditional probabilities.

c) [**1 mark**] What is the number of independent parameters that is required to describe this joint distribution?

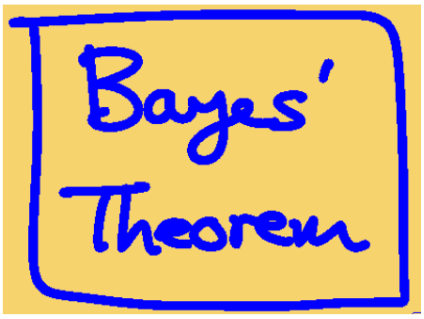
d) [**1 mark**] Assume there were no conditional independence between the variables, how many independent parameters would be required then?

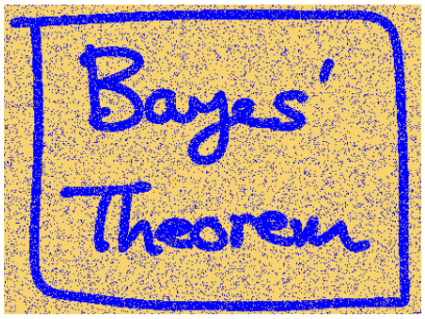
3 Markov Random Fields (5 marks)

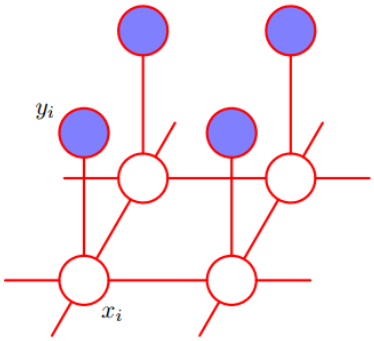
Energy Function [3 marks]

Consider the image de-noising example in the lecture. Let the observed noisy image be described by an array of binary pixel values $y_i \in \{-1, +1\}$, where the index $i = 1, ..., D$ runs over all pixels. We shall suppose that the image is obtained by taking an unknown noise-free image, described by binary pixel values $x_i \in \{-1, +1\}$ and randomly flipping the sign of pixels with some small probability. Given the noisy image, our goal is to recover the original noise-free image.

Because the noise level is small, we know that there will be a strong correlation between x_i and y_i. We also know that neighbouring pixels x_i and x_j in an image are strongly correlated. This prior knowledge can be captured using the Markov random field model.







We now take the following energy function Eq. (1), which defines a joint distribution over **x** and **y** given by Eq. (2). In Eq. (1), the $-\eta x_i y_i$ terms have the desired effect of giving a lower energy (thus higher probability) when x_i and y_i have the same sign and a higher energy when they have the opposite sign; the $-\beta x_i x_j$ terms guarantee that the energy is lower when the ground truth pixels have the same sign than when they have the opposite sign; the hx_i terms have the effect of biasing the model towards pixel values that have one particular sign in preference to the other.

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i \tag{1}$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\} \tag{2}$$

(a) [**2 mark**] Consider the use of iterated conditional modes (ICM) to minimize the energy function given by Eq. (1). Write down an expression for the change in the values of the

energy associated with the two states of a particular variable x_k, with all other variables held fixed, and show that it depends only on quantities that are local to x_k in the graph.

(b) [**1 mark**] Consider a particular case of the energy function given by Eq. (1) in which the coefficients $\beta = h = 0$. Show that the most probable configuration of the latent variables is given by $x_i = y_i$ for all i.

Undirected Graphs [2 marks]

Show that there are $2^{M(M-1)/2}$ distinct undirected graphs over a set of M distinct random variables. Draw the 8 possibilities for the case of $M = 3$.