GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

ECE 8803-GGDL Fall 2023
Problem Set #1

Assigned: 3 Sep
Due Date: 13 Sep

Please contact the TAs for clarification on the instructions in the homework assignments.

Problem 1: Data generation in low and high dimensions. In class, we identified *high dimensionality* as a critical challenge in learning generative models. In this problem, we like to train a generative model in small dimensions and then examine why it does not scale well.

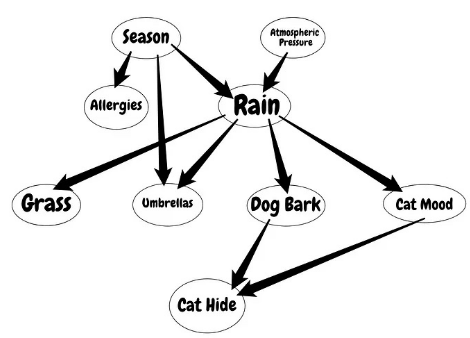
a. Consider black and white images of size 3×3. Develop a probabilistic model that generates images with a probability proportional to the number of black pixels in that image. Draw five samples and visualize them. (*HINT: use the Categorical distribution*)

b. Assume we no longer know the data generation rule and want to learn it from a handful of data points. Now, sample 200 unique images. Consider the probability of generating each training image as the label and train a Multilayer Perceptron (MLP). Use the MLP with proper normalization to develop a new probabilistic model to generate images. Draw five samples and visualize them.

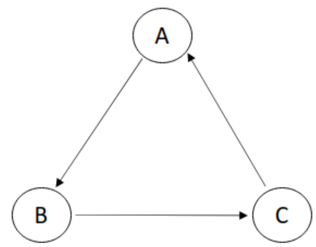
c. Now consider the same problem in 28×28 dimensions. Draw 200 random samples and train an MLP. Explain why following the same strategy to learn a generative model does not scale. Bring at least two reasons.

Problem 2: Generative modeling with independent assumption. Consider the binarized MNIST digit dataset `https://www.tensorflow.org/datasets/catalog/binarized_mnist`. Train a pixel-independent generative model by properly forming the likelihood and minimizing the maximum likelihood. Detail the mathematical steps to arrive at the model. Generate ten samples. What do you learn?

Problem 3: Cats, dogs, and DAGs. What conditional independencies can we infer from this DAG? Write your steps to find all the independencies. Which independencies do you agree with, and which do you not agree?



Problem 4: Joint probability of CGs. Consider the following directed cyclic graph. Prove why would the joint probability not be legal for this graph. *Hint: design a counter-example assuming a simple distribution for each variable (e.g., binary) and show that the attained joint distribution is invalid.*



Problem 5: NADE and MADE on MNIST. We will now compare the performance of a NADE and MADE implementation on the binarized MNIST dataset. We will use the implementation in `https://github.com/EugenHotaj/pytorch-generative`, which you will need to clone into your working directory (make sure you have the necessary Pytorch environment or follow the installation steps in the repository page).

a. Train the provided NADE model (check the `reproduce()` function provided with the model) for 30 epochs and a hidden dimension of size 300. Visualize a few samples from the trained model (you can do this either by using the `sample()` method provided with the model, or visualizing the tensorboard log after training).

b. Repeat the process with the provided MADE model (using the same parameters described above). Does one of the models perform better (generates more plausible samples) than the other? Explain why this is the case.

c. Which model would be more suitable for this particular type of data? Justify your answer using generative performance, runtime, complexity, or other aspects you consider relevant.

Problem 6: Autoregressive autoencoders. Take the autoregressive autoencoder example from the slides of lecture 3. Now assume we increase the problem dimension from 3 to 4 by adding a new variable x_4 and consider the following new ordering for the variables: x_2, x_4, x_1, x_3. Follow the steps we reviewed in class and find the new masking matrices that turn the autoencoder into an autoregressive model. Detail all your steps. *Note: there is no unique solution.*

Problem 7: MLE on GGDL. We want to estimate the size of our GGDL class. Say the students are numbered from 1 to n, where n is the number of students. We call three random students out of the class, ask for their numbers, and receive these numbers: i, j, and k. Find the maximum likelihood estimate for n.

Problem 8: MLE on the exponential family. Suppose $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ is drawn from an exponential distribution $\exp(\lambda)$. Form the maximum likelihood and estimate λ.