Homework 6

Yuhan Wang
ywang2558

Instructions: Use this latex file as a template to develop your homework. We are changing our reproducibility policy on code submissions going forward. Instead of uploading it on GitHub, please submit a separate zip file that contains your code. You will submit two files to Canvas, one is your pdf, and the other one is a zip file. Late submissions may not be accepted. You can choose any programming language (i.e. python, R, or MATLAB). Please check Piazza for updates about the homework.

1 Implementation: GAN (30 pts)

In this part, you are expected to implement GAN with MNIST dataset. We have provided a base jupyter notebook (gan-base.ipynb) for you to start with, which provides a model setup and training configurations to train GAN with MNIST dataset.

(a) Implement training loop and report learning curves and generated images in epoch 1, 50, 100. Note that drawing learning curves and visualization of images are already implemented in provided jupyter notebook. (15 pts)

Procedure 1 Training GAN, modified from ?

Input: m: real data batch size, n_z: fake data batch size
Output: Discriminator D, Generator G
for number of training iterations do
# Training discriminator
Sample minibatch of n_z noise samples $\{z^{(1)}, z^{(2)}, \cdots, z^{(n_z)}\}$ from noise prior $p_g(z)$
Sample minibatch of $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d}\Big(\frac{1}{m}\sum_{i=1}^{m}\log D(x^{(i)}) + \frac{1}{n_z}\sum_{i=1}^{n_z}\log(1 - D(G(z^{(i)})))\Big)$$

Training generator
Sample minibatch of n_z noise samples $\{z^{(1)}, z^{(2)}, \cdots, z^{(n_z)}\}$ from noise prior $p_g(z)$
Update the generator by ascending its stochastic gradient:

$$\nabla_{\theta_g}\frac{1}{n_z}\sum_{i=1}^{n_z}\log D(G(z^{(i)}))\Big)$$

end for
# The gradient-based updates can use any standard gradient-based learning rule. In the base code, we are using Adam optimizer (?)

Expected results are as follows.

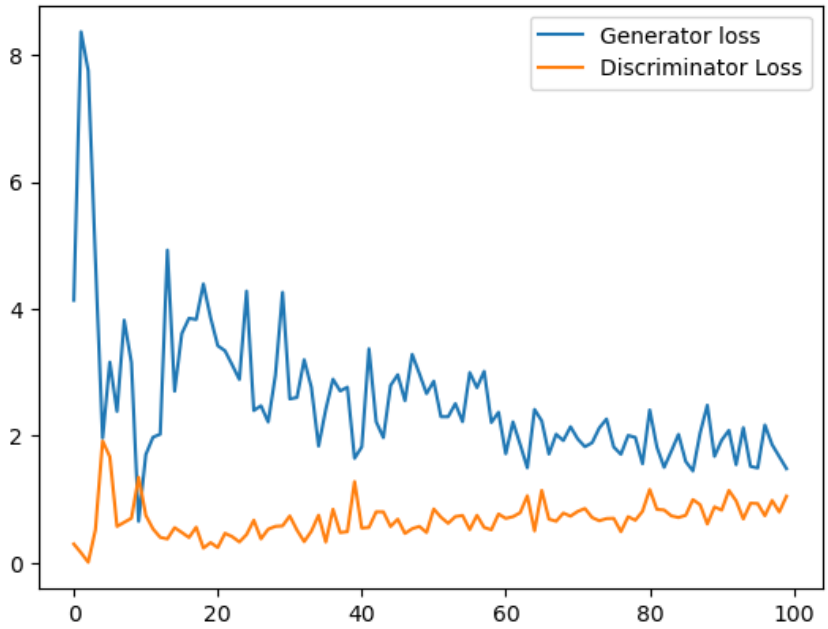
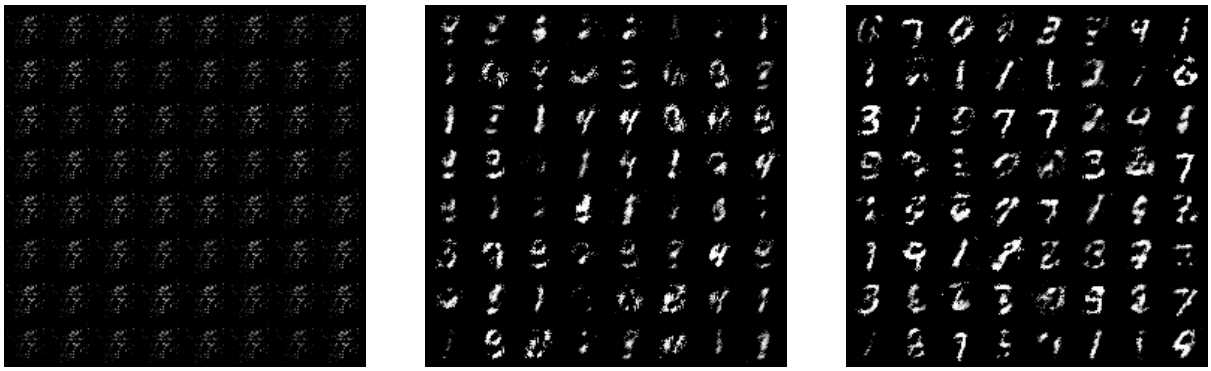


Figure 1: Learning curve



(a) epoch 1 (b) epoch 50 (c) epoch 100

Figure 2: Generated images by G

(b) Replace the generator update rule as the original one in the slide,
"Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{n_z} \sum_{i=1}^{n_z} \log(1 - D(G(z^{(i)})))$$

" , and report learning curves and generated images in epoch 1, 50, 100. Compare the result with (a). Note that it may not work. If training does not work, explain why it doesn't work. (10 pts)

This objective function can not work directly since the probability generated by discriminator is too small for gradient optimization. So I take the log of the probability, divide it by 1000 then take the exponential. The result is is shown in Fig.3 and Fig.4.

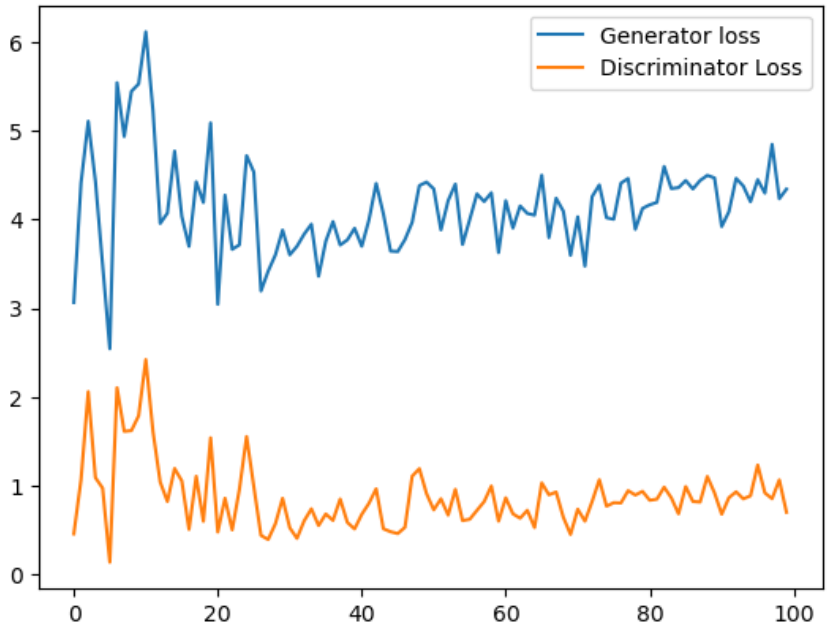
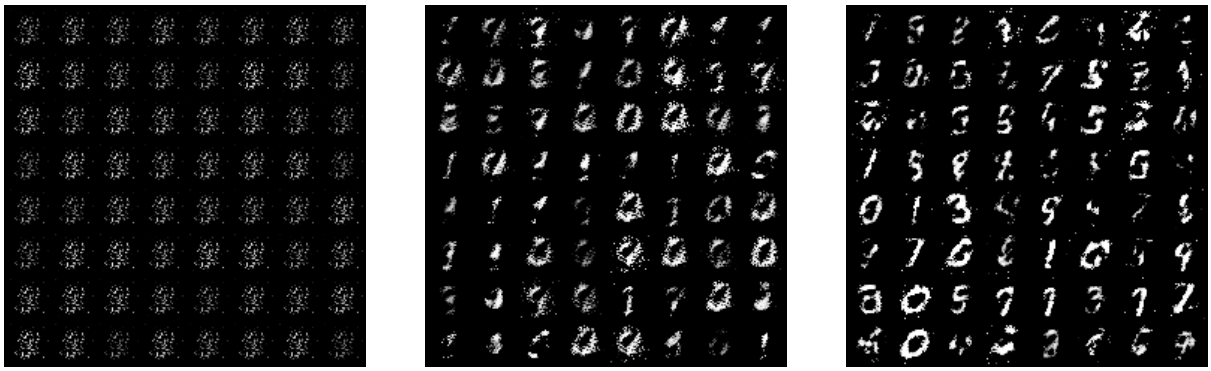


Figure 3: Learning curve



(a) epoch 1 (b) epoch 50 (c) epoch 100

Figure 4: Generated images by G

(c) Except the method that we used in (a), how can we improve training for GAN? Implement that and report your setup, learning curves, and generated images in epoch 1, 50, 100. (5 pts)

By increasing the number of discriminator and generator updates per generator update, we can improve the training for GAN. In the base code, we have set the number of discriminator updates per generator update to 1. We can increase this number to 5, and the result is shown in Fig.5 and Fig.6.

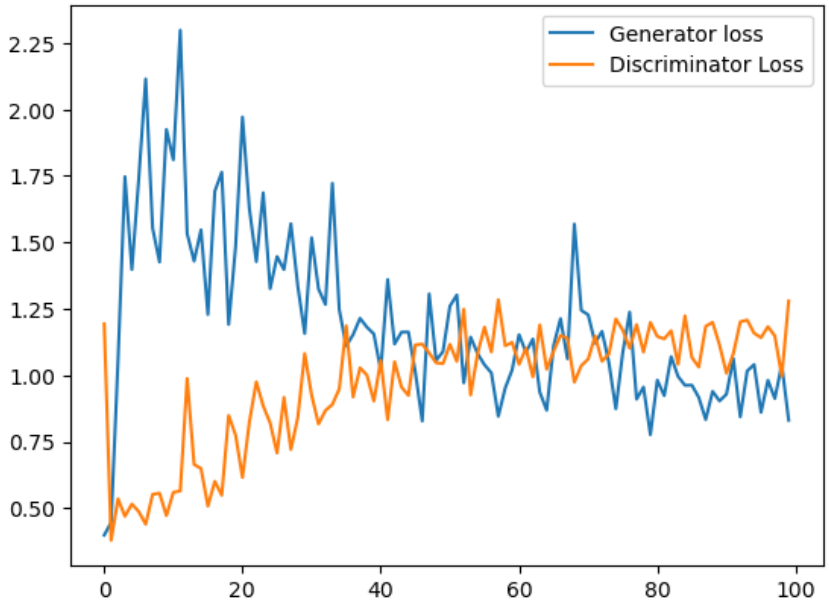
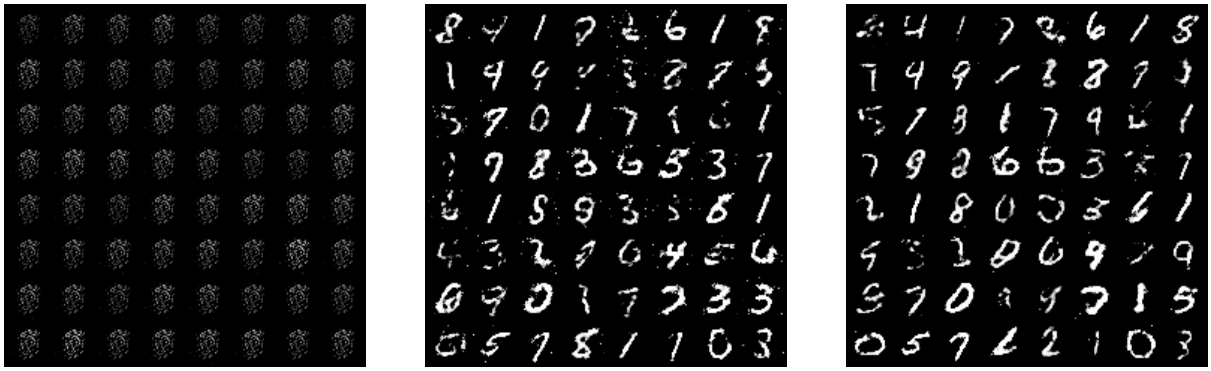


Figure 5: Learning curve of q1.3



(a) epoch 1 (b) epoch 50 (c) epoch 100

Figure 6: Generated images of q1.3

2 Review change of variables in probability density functions [25 pts]

In Flow based generative model, we have seen $p_\theta(x) = p(f_\theta(x))|\frac{\partial f_\theta(x)}{\partial x}|$. As a hands-on (fixed parameter) example, consider the following setting.
Let X and Y be independent, standard normal random variables. Consider the transformation $U = X+Y$ and $V = X-Y$. In the notation used above, $U = g_1(X,Y)$ where $g_1(X,Y)$ where $g_1(x,y) = x+y$ and $V = g_2(X,Y)$ where $g_2(x,y) = x-y$. The joint pdf of X and Y is $f_{X,Y} = (2\pi)^{-1}exp(-x^2/2)exp(-y^2/2), -\infty < x < \infty, -\infty < y < \infty$. Then, we can determine u, v values by x, y, i.e. $\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$.

(a) Compute Jacobian matrix

$$J = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

(5 pts)

$$J = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{\partial (u+v)/2}{\partial u} & \frac{\partial (u+v)/2}{\partial v} \\ \frac{\partial (u-v)/2}{\partial u} & \frac{\partial (u-v)/2}{\partial v} \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}$$

(b) (Forward) Show that the joint pdf of U, V is

$$f_{U,V}(u,v) = \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-u^2/4)\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-v^2/4)\right)$$

(10 pts)

(Hint: $f_{U,V}(u,v) = f_{X,Y}(?,?)|det(J)|$)

$|det(J)| = \frac{1}{2}$, then

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(x,y)|det(J)| \\ &= \left(\frac{1}{\sqrt{2\pi}} exp(-x^2/2)\right)\left(\frac{1}{\sqrt{2\pi}} exp(-y^2/2)\right)\frac{1}{2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-(u+v)^2/8)\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-(u-v)^2/8)\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-u^2/4)\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-v^2/4)\right) \end{aligned}$$

(c) (Inverse) Check whether the following equation holds or not.

$$f_{X,Y}(x,y) = f_{U,V}(x+y, x-y)|det(J)^{-1}|$$

(10 pts)

Yes, it holds as follows:

$$\begin{aligned} f_{X,Y}(x,y) &= f_{U,V}(u,v)|det(J)^{-1}| \\ &= \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-u^2/4)\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} exp(-v^2/4)\right)2 \\ &= \left(\frac{1}{\sqrt{2\pi}} exp(-(x+y)^2/4)\right)\left(\frac{1}{\sqrt{2\pi}} exp(-(x-y)^2/4)\right) \\ &= \left(\frac{1}{\sqrt{2\pi}} exp(-x^2/2)\right)\left(\frac{1}{\sqrt{2\pi}} exp(-y^2/2)\right) \end{aligned}$$

3 Directed Graphical Model [20 points]

Consider the directed graphical model (aka Bayesian network) in Figure 7.

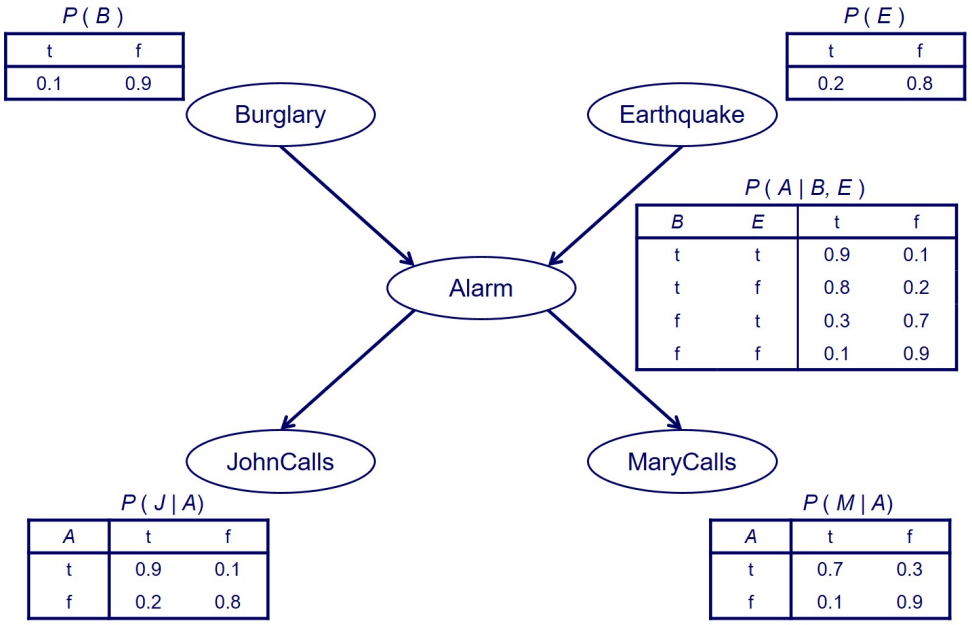


Figure 7: A Bayesian Network example.

Compute $P(B = t \mid E = f, J = t, M = t)$ and $P(B = t \mid E = t, J = t, M = t)$. (10 points for each) These are the conditional probabilities of a burglar in your house (yikes!) when both of your neighbors John and Mary call you and say they hear an alarm in your house, but without or with an earthquake also going on in that area (what a busy day), respectively.

$$P(B = t \mid E = f, J = t, M = t) = \frac{P(B = t, E = f, J = t, M = t)}{P(E = f, J = t, M = t)}$$

$$\begin{aligned} P(B = t, E = f, J = t, M = t) &= P(B = t)P(E = f)P(A = t \mid B = t, E = f)P(J = t \mid A = t)P(M = t \mid A = t) + \\ &\quad P(B = t)P(E = f)P(A = f \mid B = t, E = f)P(J = t \mid A = f)P(M = t \mid A = f) \\ &= 0.1 \times 0.8 \times 0.8 \times 0.9 \times 0.7 + 0.1 \times 0.8 \times 0.2 \times 0.2 \times 0.1 \\ &= 0.04032 + 0.00032 = 0.04064 \end{aligned}$$

$$\begin{aligned} P(E = f, J = t, M = t) &= P(B = t)P(E = f)P(A = t \mid B = t, E = f)P(J = t \mid A = t)P(M = t \mid A = t) + \\ &\quad P(B = t)P(E = f)P(A = f \mid B = t, E = f)P(J = t \mid A = f)P(M = t \mid A = f) + \\ &\quad P(B = f)P(E = f)P(A = t \mid B = f, E = f)P(J = t \mid A = t)P(M = t \mid A = t) + \\ &\quad P(B = f)P(E = f)P(A = f \mid B = f, E = f)P(J = t \mid A = f)P(M = t \mid A = f) \\ &= 0.1 \times 0.8 \times 0.8 \times 0.9 \times 0.7 + 0.1 \times 0.8 \times 0.2 \times 0.2 \times 0.1 + \\ &\quad 0.9 \times 0.8 \times 0.1 \times 0.9 \times 0.7 + 0.9 \times 0.8 \times 0.9 \times 0.2 \times 0.1 \\ &= 0.04032 + 0.00032 + 0.04536 + 0.01296 = 0.09896 \end{aligned}$$

$$P(B = t \mid E = f, J = t, M = t) = \frac{0.04064}{0.09896} = 0.4107$$

$$P(B=t \mid E=t, J=t, M=t) = \frac{P(B=t, E=t, J=t, M=t)}{P(E=t, J=t, M=t)}$$

$$\begin{aligned}
P(B=t, E=t, J=t, M=t) &= P(B=t)P(E=t)P(A=t \mid B=t, E=t)P(J=t \mid A=t)P(M=t \mid A=t) + \\
&\quad P(B=t)P(E=t)P(A=f \mid B=t, E=t)P(J=t \mid A=f)P(M=t \mid A=f) \\
&= 0.1 \times 0.2 \times 0.9 \times 0.9 \times 0.7 + 0.1 \times 0.2 \times 0.1 \times 0.2 \times 0.1 \\
&= 0.01134 + 0.00004 = 0.01138 \\
P(E=t, J=t, M=t) &= P(B=t)P(E=t)P(A=t \mid B=t, E=t)P(J=t \mid A=t)P(M=t \mid A=t) + \\
&\quad P(B=t)P(E=t)P(A=f \mid B=t, E=t)P(J=t \mid A=f)P(M=t \mid A=f) + \\
&\quad P(B=f)P(E=t)P(A=t \mid B=f, E=t)P(J=t \mid A=t)P(M=t \mid A=t) + \\
&\quad P(B=f)P(E=t)P(A=f \mid B=f, E=t)P(J=t \mid A=f)P(M=t \mid A=f) \\
&= 0.1 \times 0.2 \times 0.9 \times 0.9 \times 0.7 + 0.1 \times 0.2 \times 0.1 \times 0.2 \times 0.1 + \\
&\quad 0.9 \times 0.2 \times 0.3 \times 0.9 \times 0.7 + 0.9 \times 0.2 \times 0.7 \times 0.2 \times 0.1 \\
&= 0.01134 + 0.00004 + 0.03402 + 0.00252 = 0.04854
\end{aligned}$$

$$P(B=t \mid E=t, J=t, M=t) = \frac{0.01138}{0.04854} = 0.2344$$

4 Chow-Liu Algorithm [25 pts]

Suppose we wish to construct a directed graphical model for 3 features X, Y, and Z using the Chow-Liu algorithm. We are given data from 100 independent experiments where each feature is binary and takes value T or F. Below is a table summarizing the observations of the experiment:

X	Y	Z	Count
T	T	T	36
T	T	F	4
T	F	T	2
T	F	F	8
F	T	T	9
F	T	F	1
F	F	T	8
F	F	F	32

1. Compute the mutual information $I(X,Y)$ based on the frequencies observed in the data. (5 pts)

$$\begin{aligned}
I(X,Y) &= P(X=T,Y=T)log\frac{P(X=T,Y=T)}{P(X=T)P(Y=T)} + P(X=T,Y=F)log\frac{P(X=T,Y=F)}{P(X=T)P(Y=F)} + \\
&\quad P(X=F,Y=T)log\frac{P(X=F,Y=T)}{P(X=F)P(Y=T)} + P(X=F,Y=F)log\frac{P(X=F,Y=F)}{P(X=F)P(Y=F)} \\
&= 0.4 \times log\frac{0.4}{0.5 \times 0.5} + 0.1 \times log\frac{0.1}{0.5 \times 0.5} + 0.1 \times log\frac{0.1}{0.5 \times 0.5} + 0.4 \times log\frac{0.4}{0.5 \times 0.5} \\
&= 0.2781
\end{aligned}$$

2. Compute the mutual information $I(X, Z)$ based on the frequencies observed in the data. (5 pts)

$$
\begin{aligned}
I(X,Z) &= P(X=T,Z=T)log\frac{P(X=T,Z=T)}{P(X=T)P(Z=T)} + P(X=T,Z=F)log\frac{P(X=T,Z=F)}{P(X=T)P(Z=F)} + \\
&\quad P(X=F,Z=T)log\frac{P(X=F,Z=T)}{P(X=F)P(Z=T)} + P(X=F,Z=F)log\frac{P(X=F,Z=F)}{P(X=F)P(Z=F)} \\
&= 0.38 \times log\frac{0.38}{0.5 \times 0.55} + 0.12 \times log\frac{0.12}{0.5 \times 0.45} + 0.17 \times log\frac{0.17}{0.5 \times 0.55} + 0.33 \times log\frac{0.33}{0.5 \times 0.45} \\
&= 0.1328
\end{aligned}
$$

3. Compute the mutual information $I(Z, Y)$ based on the frequencies observed in the data. (5 pts)

$$
\begin{aligned}
I(Z,Y) &= P(Z=T,Y=T)log\frac{P(Z=T,Y=T)}{P(Z=T)P(Y=T)} + P(Z=T,Y=F)log\frac{P(Z=T,Y=F)}{P(Z=T)P(Y=F)} + \\
&\quad P(Z=F,Y=T)log\frac{P(Z=F,Y=T)}{P(Z=F)P(Y=T)} + P(Z=F,Y=F)log\frac{P(Z=F,Y=F)}{P(Z=F)P(Y=F)} \\
&= 0.45 \times log\frac{0.45}{0.55 \times 0.5} + 0.1 \times log\frac{0.1}{0.55 \times 0.5} + 0.05 \times log\frac{0.05}{0.45 \times 0.5} + 0.4 \times log\frac{0.4}{0.45 \times 0.5} \\
&= 0.3973
\end{aligned}
$$

4. Which undirected edges will be selected by the Chow-Liu algorithm as the maximum spanning tree? (5 pts)

The edges selected by the Chow-Liu algorithm are:
(X, Y) and (Z, Y)

5. Root your tree at node X, assign directions to the selected edges. (5 pts)

The edges selected by the Chow-Liu algorithm are:
(X, Y) and (Z, Y)
Rooting the tree at node X, we get:
(X, Y) and (Y, Z)