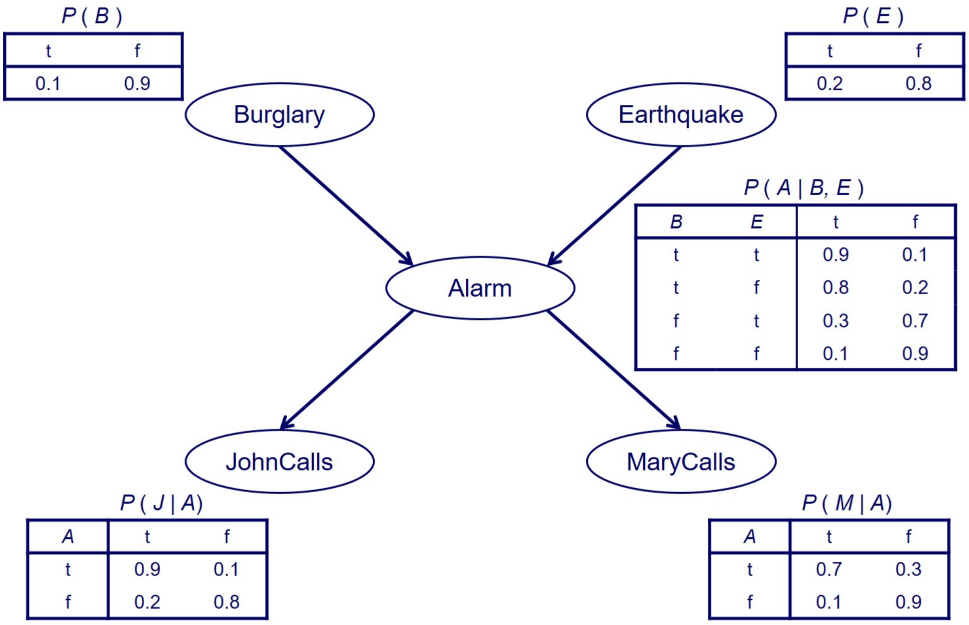
HOMEWORK 7

HAMSALEKHA PREMKUMAR

Instructions: Although this is a programming homework, you only need to hand in a pdf answer file. There is no need to submit the latex source or any code. You can choose any programming language, as long as you implement the algorithm from scratch.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Please check Piazza for updates about the homework.

1 Directed Graphical Model [30 points]

Consider the directed graphical model (aka Bayesian network) in Figure 1.



P (B)

t	f
0.1	0.9

P (E)

t	f
0.2	0.8

P (A | B, E)

B	E	t	f
t	t	0.9	0.1
t	f	0.8	0.2
f	t	0.3	0.7
f	f	0.1	0.9

P (J | A)

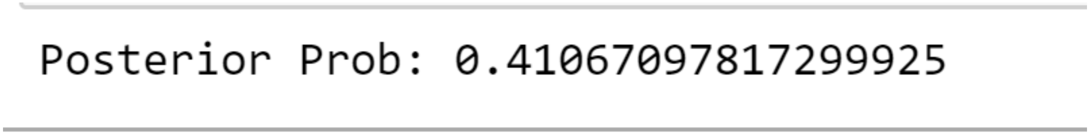
A	t	f
t	0.9	0.1
f	0.2	0.8

P (M | A)

A	t	f
t	0.7	0.3
f	0.1	0.9

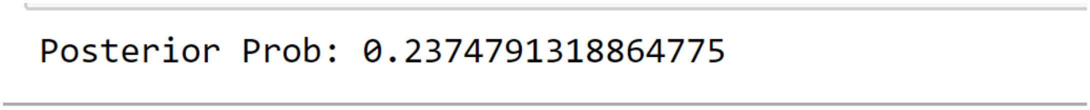
Figure 1: A Bayesian Network example.

Compute $P(B = t \mid E = f, J = t, M = t)$ and $P(B = t \mid E = t, J = t, M = t)$. 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.



```
Posterior Prob: 0.41067097817299925
```

Fig: $P(B = t \mid E = f, J = t, M = t)$



```
Posterior Prob: 0.2374791318864775
```

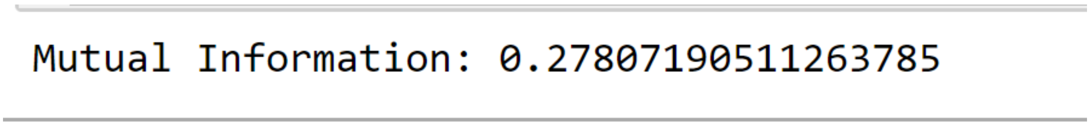
Fig: $P(B = t \mid E = t, J = t, M = t)$

2 Chow-Liu Algorithm [40 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

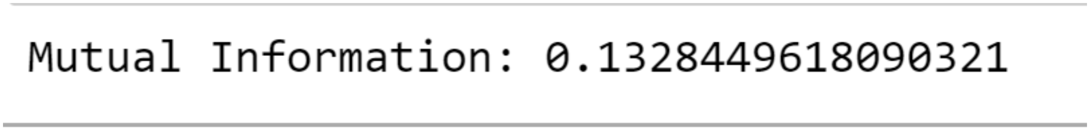
1. Compute the mutual information $I(X, Y)$ based on the frequencies observed in the data.



```
Mutual Information: 0.2780719051126385
```

Fig: $I(X, Y)$

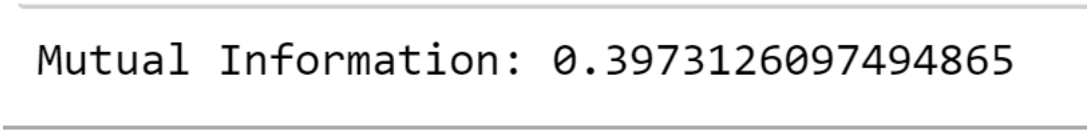
2. Compute the mutual information $I(X, Z)$ based on the frequencies observed in the data.



```
Mutual Information: 0.1328449618090321
```

Fig: $I(X, Z)$

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



```
Mutual Information: 0.3973126097494865
```

Fig: $I(Z, Y)$

4. Which undirected edges will be selected by the Chow-Liu algorithm as the maximum spanning tree?
   Chow-Lui algorithm sorts the edges in descending order of weights, picks those edges that do not form a cycle as the edges of a tree. Thus edges (Z,Y) and (X,Y) are selected for the maximum spanning tree.

5. Root your tree at node X, assign directions to the selected edges.
   Arbitrarily choosing direction we get the following tree with X and the root node, Y as child of X and Z as child of Y.
   X → Y → Z

3 Kernel SVM [30 points]

Consider the following kernel function defined over $z, z' \in Z$:

$$k(z, z') = \begin{cases} 1 & \text{if } z = z', \\ 0 & \text{otherwise.} \end{cases}$$

1. Prove that for any integer $m > 0$, any $z_1, \ldots, z_m \in Z$, the $m \times m$ kernel matrix $K = [K_{ij}]$ is positive semi-definite, where $K_{ij} = k(z_i, z_j)$ for $i, j = 1 \ldots m$. (Let us assume that for $i \neq j$, we have $z_i \neq z_j$.) Hint: An $m \times m$ matrix K is positive semi-definite if $\forall v \in \mathbb{R}^m, v^\top K v \geq 0$.

   From the above conditions, $K_{ij} = k(z_i, z_j)$, K turns out to be an $m \times m$ Identity matrix. $K = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$

   Consider an arbitrary vector,$v \in \mathbb{R}^m$, $v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$

   $$v^\top K v = \begin{bmatrix} v_1 & v_2 & v_3 & \ldots & v_m \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$
   $$= v_1^2 + v_2^2 + v_3^2 + ... + v_m^2$$
   $$\geq 0$$

   for any value of of v_i. Thus the matrix $K_{m \times m}$ is positive semi-definite.

2. Given a training set $(z_1, y_1), \ldots, (z_n, y_n)$ with binary labels, the dual SVM problem with the above kernel k will have parameters $\alpha_1, \ldots, \alpha_n, b \in \mathbb{R}$. (Assume that for $i \neq j$, we have $z_i \neq z_j$.) The predictor for input z takes the form

   $$f(z) = \sum_{i=1}^{n} \alpha_i y_i k(z_i, z) + b.$$

   Recall the label prediction is $\text{sgn}(f(z))$. Prove that there exists $\alpha_1, \ldots, \alpha_n, b$ such that f correctly separates the training set.

   In other words, k induces a feature space rich enough such that in it any training set can be linearly separated.

   f correctly separates the training set when $\text{sgn}(f(z_i)) = y_i$ for all $i = 1, 2, ...n$

   ie., $\text{sgn}(f(z_k)) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i k(z_i, z_k) + b) = y_k$, $k = 1, 2...n$

   From 1, we know that

   $$K = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

   Thus,for some α and b, $\text{sgn}(f(z_1)) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i k(z_i, z_1) + b) = \text{sgn}(\alpha_1 y_1 + b) = y_1$
   $\text{sgn}(f(z_2)) = \text{sgn}(\alpha_2 y_2 + b) = y_2$
   .
   .
   .
   $\text{sgn}(f(z_n)) = \text{sgn}(\alpha_n y_n + b) = y_n$

   Observing this system of equations, it can be concluded that f correctly separates the training set if $\alpha_i y_i + b \geq 0$ as $\text{sgn}(0) = 0$ (classifies as class 0) and $\text{sgn}(positivenumber) = 1$ (classifies as class 1) and one of the many possible solutions is if $\alpha_i > 0$ and $b = 0$.

3. How does that f predict input z that is not in the training set?
   When input z is not in the training set, $k(z_i, z) = 0$. Hence $\text{sgn}(f(z)) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i k(z_i, z) + b) = \text{sgn}(b)$. Since α_i are 0,essentially, the features of the test data are irrelevant and this model arbitrarily classifies the test data based on $\text{sgn}(b)$. This shows that the model knows nothing outside the train set.

Comment: One useful property of kernel functions is that the input space Z does not need to be a vector space; in other words, z does not need to be a feature vector. For all we know, Z can be turkeys in the world. As long as we can compute $k(z, z')$, kernel SVM works on turkeys.

4 Extra Credit: Kernel functions over discrete space [10 points]

Kernel functions can be defined over objects as diverse as graphs, sets, strings, and text documents. Consider, for instance, a fixed set D and define a nonvectorial space consisting of all possible subsets of this set D. If A_1 and A_2 are two such subsets then one simple choice of kernel would be

$$k(A_1, A_2) = 2^{|A_1 \cap A_2|}$$

where $A_1 \cap A_2$ denotes the intersection of sets A_1 and A_2, and $|A|$ denotes the size of A. Show that this is a valid kernel function, by showing that it corresponds to an inner product in a feature space.
Given A_1 and A_2 are subsets of D, to prove: $k(A_1, A_2) = 2^{|A_1 \cap A_2|} = \phi(A_1)^T \phi(A_2)$
Since A_1 and A_2 are subsets of D, $\phi(A)$ is a $2^{|D|} \times 1$ vector such that each element is a certain subset of D. And hence $\phi(A_i)$ has $2^{|D|}$ elements. For simplicity of representation, each element of the column vector has 1 in those positions corresponding to different subsets of D and 0s otherwise. Then, $\phi(A_1)^T \phi(A_2)$ iterates over each subset of D(given by A) and gives the count of subsets common to both A_1 and A_2 which is equivalent to $2^{|A_1 \cap A_2|}$. Thus, $k(A_1, A_2) = 2^{|A_1 \cap A_2|} = \phi(A_1)^T \phi(A_2)$.And hence a valid kernel.

5 Extra Credit: Support Vector Machines [10 points]

Given data $\{(x_i, y_i), 1 \leq i \leq n\}$, the (hard margin) SVM objective is

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}\|w\|_2^2 \\ &\text{s.t. } \ y_i(w^\top x_i + b) \geq 1 (\forall i). \end{aligned}$$

The dual is

$$\begin{aligned} &\max_{\alpha} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ &\text{s.t. } \ \alpha_i \geq 0 (\forall i), \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

Suppose the optimal solution for the dual is $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$, and the optimal solution for the primal is (w^*, b^*). Show that the margin

$$\gamma = \min_i \frac{y_i((w^*)^\top x_i + b^*)}{\|w^*\|_2}$$

satisfies

$$\frac{1}{\gamma^2} = \sum_{i=1}^{n} \alpha_i^*.$$

Hint: use the KKT conditions.
Optimal (w^*, b^*) results in min margin =1. Hence we take $y_i((w^*)^\top x_i + b^*) = 1$
Thus

$$\gamma = \frac{1}{\|w^*\|_2}$$

And from KKT condtions, $w = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.

$(\gamma)^2 = \frac{1}{\sum_{i=1}^{n} \alpha_i^* y_i x_i}$
Applying KKT conditions along with $y_i((w^*)^\top x_i + b^*) = 1$ we have $\sum_{i=1}^{n} \alpha_i^* = \frac{1}{(\gamma)^2}$