CS 228 Winter 2018 Homework 1

SUNet ID: 05794739
Name: Anass Belcaid
Collaborators:
Late Days: 2

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1: Probability theory (4 points)

The doctor has bad news and good news for X. The bad news is that X tested positive for a serious disease and the test is 99% accurate. The good news is that is a rare disease, striking at only one in 10,000 people.

- What it is a good news that the disease is rare?
- What are the chances that X has actually the disease?

(a) The fact that the disease is rare makes the prior $P(D = 1) = 10^{-5}$ makes the probability of having the disease so small even if the test is positive.

(b) Let denote by D the probability of having the disease, hence, we have $P(D = 1) = 10^{-5}$. Also we will denote the conditional probability $P(D|T)$ of having the disease D given the the result of the test T.

Hence the query probability is :

$$\begin{aligned} P(D|T) &= \frac{P(T|D)\ P(D)}{P(T)} && (1)\\ &= \frac{P(T|d=1)P(D=1)}{P(T|D=1)P(D=1) + P(T|D=0)P(D=0)} && (2)\\ &= \frac{0.99 \times 0.00001}{0.99 \times 0.00001 + 0.01 \times 0.99999} && (3)\\ &\approx 0.98\% && (4) \end{aligned}$$

Therefore even if the test is 99%, the probability of X to have to disease is still very **low**.

Problem 2: Review of dynamic programming (7 points)

Suppose you have a probability distribution P over the variables $X_1, X_2, \ldots, X_n$ which all take values on the set $\mathcal{S} = \{v_1, \ldots, v_m\}$ where v_j are some distinct variables (digits or letters). Suppose that P satisfies the *Markov assumption* for all $i \geq 2$ we have:

$$P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|x_{i-1}) \tag{5}$$

In other word P factorizes as

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_2|x_1)\ldots P(x_n|x_{n-1}) \tag{6}$$

for each $i \geq 2$, you are given the factor $P(x_i|x_{i-1})$ as an $m \times m$ table and you are given the table $P(x_1 = v)$ for each $v \in \mathcal{S}$.

- Give an $\mathcal{O}(m^2 n)$ algorithm for solving the problem

$$\max_{x1,\ldots,x_m \in \mathcal{S}^n} P(x_1)P(x_2|x_1)\ldots P(x_n|x_{n-1}) \tag{7}$$

(a) In order to solve the problem in Eq.7, we propose the following *reduction* to the problem

$$\max_{x1,\ldots,x_m \in \mathcal{S}^n} P(x_1,\ldots,x_n) = \max_{x_n,x_{n-1}} P(x_n|x_{n-1}) \max_{x_1,\ldots,x_{n-1}} P(x_1)\ldots P(x_{n-1}|x_{n-2}) \tag{8}$$

- The first maximization could be solved in quadratic time $\mathcal{O}(m^2)$, since we need to check all the possible values for x_n and x_{n-1}.
- By doing so, we simplified the problem from computing the **maximum** over n variable to a maximum of $n-1$ variables. hence, we will repeat the simplification until we obtain the unique factor $P(x_1)$.

Algorithm 1 Dynamic programming to solve the maximization problem

Require: factors $P(x_i|x_{i-1})$
  initialize $\mathbf{M} \in \mathbf{R}^m$ with $P(x_1 = v \in \mathcal{S})$
  for i = 1 **to** n **do**
    for j = 1 **to** m **do**
      max_value $= -\infty$
      for k = 1 **to** m **do**
        max_value = max(max_value, $M[j] \times P(X_i = v_k|V_{i-1} = v_j)$)
      end for
    end for
  end for

1 Problem 3: Bayesian Network (6 points)

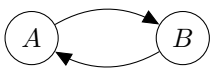
Let's try to relax the definition of *Bayesian network* by removing the assumption that the directed graph is acyclic. Suppose that we have a directed graph $G = (V, E)$ and a discrete random variables $X_1, \ldots, X_n$ and define:

$$f(x_1,\ldots,x_n) = \prod_{v \in V} f_v(x_v|x_{pa}(v)) \tag{9}$$

where x_{pa} refers to the parents of of the variable X_v in G and f_v specifies the distribution over the X_v for every assignment of the parents. We recall that Eq.9 is precisely the definition of the joint probability associated with Bayesian network G, where f_v are the conditional probabilities.

- Show that if the graph has a directed cycle, f do no longer define a probability distribution.
- In particular, give an example of a cyclic graph G and a distribution f_v that lead to improper probabilities.

(a) Let's consider the simple bayesian network



with the following tables $P(A|B) = \begin{bmatrix} & b_0 & b_1 \\ a_0 & 1 & 0 \\ a_1 & 0 & 1 \end{bmatrix}$ and $P(B|A) = \begin{bmatrix} & a_0 & a_1 \\ b_0 & 0.7 & 0.5 \\ b_1 & 0.3 & 0.5 \end{bmatrix}$.

With this choice we get

a, b	$P(a, b)$
(a_0, b_0)	0.7
(a_0, b_1)	0
(a_1, b_0)	0
(a_1, b_1)	0.5

we clearly is not a probability distribution since $\sum_{a,b} P(A, B) \neq 1$.

2 Problem 4: Conditional Independence (12 points)

The question investigates the way in which conditional independence relationship affect the amount of information needed for probabilistic graphical calculations. Let α and β and γ be three random variables.

- (6 points) Suppose we wish to calculate $P(\alpha|\beta, \gamma)$ and we have no conditional independence information. Which of the following set of number is sufficient for the calculations?

  1. $P(\beta, \gamma)$, $P(\alpha)$, $P(\beta|\alpha)$ and $P(\gamma|\alpha)$.
  2. $P(\beta, \gamma)$, $P(\alpha)$ and $P(\beta, \gamma|\alpha)$
  3. $P(\beta|\gamma)$, $P(\gamma|\alpha)$ and $P(\alpha)$

  for each case either give the method to calculate the probability or why it's not sufficient to compute the posterior.

- (6 points) Suppose we know β and γ are independent given α. Now which one of preceding probabilities is sufficient.

(a) For the first case the probability $P(\gamma, \beta)$ is the probability of the conditioned event. Hence, we need to express the probability $P(\alpha, \beta, \gamma)$ with the remaining expressions. we have $P(\alpha)$ and $P(\beta|\alpha)$ the third one need to **forcefully** to be conditioned on α and β.

(b) The second one could be used to get the posterior using *Bayes* rule

$$P(\alpha|\beta, \gamma) = \frac{P(\alpha)P(\beta, \gamma|\alpha)}{P(\beta, \gamma)} \tag{10}$$

(c) For the third one, we can no longer calculate the probability of the conditioned event $P(\beta, \gamma)$ by the given expressions.

(a) If we have the information β and γ are independent given α we could use the entities in the **first** proposition by

$$P(\alpha|\beta, \gamma) = \frac{P(\alpha)P(\gamma|\alpha)P(\beta|\alpha)}{P(\gamma, \beta)} \tag{11}$$

3 Bayesian network (AD exercice 4.1): (5 points)

(a) From the graph, the nodes A and B are independents.

$$P(A=0,B=0) = P(A=0)P(B=0) = 0.8 \times 0.3 = 0.24$$

for the second P(E=1—A = 1), since there is no active **trail** from A to E the two variables are independents. Hence $P(E=1|A=1) = P(E=1)$.

$$P(E=1) = \sum_B P(E=1|B)P(B) = 0.7 \times 0.1 + 0.3 \times 0.9 = 0.34$$

(b) Are A and E D-separable given E and H. No since there is an active trail $A-C-F-H-E$ leading from A to E. The trail is active since it contains the $V-$ structure $F-H-E$.

(c) Are G and E D-separable given D. Yes since there is no active path from G to E or vise-versa

(d) Are A, B and G, H D-separable given F. **False** since there is a clear path $B-E-H$.

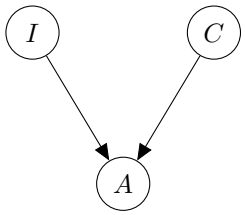
4 Bayesian networks: Explaining away (7 points)

You want to model the admission process of Farm university. Students are admitted based on their Creativity (C) and Intelligence (I). You decide to model them as continuous random variables, and your data suggests that both are uniformly distributed in $[0,1]$, and are independent of each other. Formally $I \sim \mathcal{U}([0,1])$, $C \sim \mathcal{U}([0,1])$ and $I \perp C$.

Being very prestigious the school only admits student such as $I + C \geq 1.5$

1. (1 points) What is the expected creativity score of a student?
2. (2 points) What is the expected creativity score of an admitted student?
3. (2 points) What is the expected creativity score of a student with $I = 0.95$.
4. (2 points) What is the expected creativity score of an admitted student with intelligence $I = 0.95$? How does this score compares to the question 3?

(a) Let's draw the graphical model for the three variables



Since the single variable $I \sim \mathcal{U}(0,1)$, its expectation $E(I) = 0.5$.

(b) The expected creativity of an admitted student is :

$$E(C|A=1) = \int_0^1 cP(c|A=1)dc \tag{12}$$

Or We need to compute the probability of the event A.

$$P(A) = \int_c \int_i f_I \, f_C di \, dc \tag{13}$$

$$= \int_{.5}^{1} di \int_{1.5-i}^{1} dc \tag{14}$$

$$= = \frac{1}{8} \tag{15}$$

$$\tag{16}$$

and the conditioned probability probability density $f_{c|A}$ is given by:

$$f_{c|A} = \begin{cases} \frac{1}{8} & \text{if } A \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Finally the expected creativity $E(c|A)$ is given by:

$$E(C|A) = \int_{.5}^{1} di \int_{1.5-i}^{1} cdc = \frac{5}{6} \tag{18}$$

(c) Given the graphical model, the variable **C** is independent of the intelligence. Therefore:

$$E(C|I = 0.95) = E(C) = 0.5 \tag{19}$$

(d) In order to compute the expected creativity of an admitted student with Intelligence $I = 0.95$, We need to compute the probability of the event A and $I = 0.85$.

$$P(A, I = 0.95) = \int_{c=0.55}^{1} dc = 0.45 \tag{20}$$

Given that, the conditioned expectation $E(C|A, I = 0.95)$ is given by:

$$E(C|A, I = 0.95) = \int_{.55}^{1} \frac{c}{0.45} dc = \mathbf{0.77} \tag{21}$$

The expected creativity did decrease indicating the dependence of the two variables I and C given we observed the common effect A.

5 Bayesian Networks (3.11 from Kroller)

1. (8 points) Consider the Burglary alarm given in figure 1. Construct a Bayesian Network over all the node **except** the *Alarm* That is a minimal *I-map* for the marginal distribution over the remaining variables.[1].

   Let's redraw the network by taking the first letter of each variable ans making the variable `Alarm` oberved.

[1] Be sure to get all the dependencies from the original graph

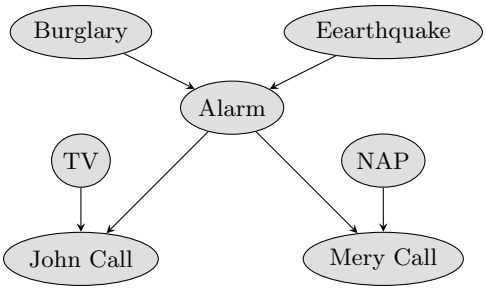
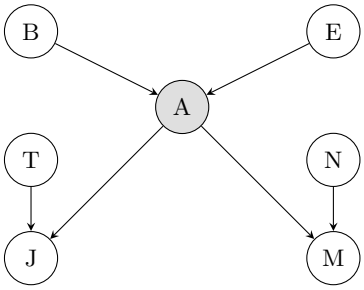
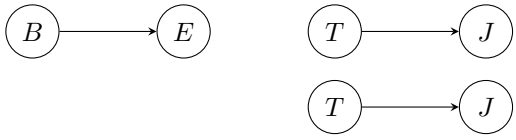


Figure 1: Bayesian network from Kroller 3.11



We construct a **topological sorring** of the graph and then for each node X_i, we chooose the minimal set U that eclipse the X_i for the rest of the other variables $\big(X_i|\{X_0,\ldots,X_{i-1}\}-U|U\big)$.



2. (8 points) Generalize this procedure to an arbitrary network. More precisely, assume we are given a network BN, an ordering $X_1,\ldots,X_n$ that is consistent with the ordering of the variables in BN and a node X_i to be removed. Specify a network BN1 is minimal *I-map* for the remaining variables. Similar to the procedure in **Koller 3.4(p 79)**, we will start with the vraiables X_1 eand for each variable X_i, we will choose the minitmal set $\mathcal{D}$ such as

$$X_i \perp \imath\{X_1,\ldots,X_{i-1}\}-\mathcal{D}\,|\mathcal{D},X_r\} \tag{22}$$

Once we computed $\mathcal{D}$, we chose the set of its nodes as the parents of X_i. Therefore, we add an arc $X_j \to X_i$ for each $X_j \in \mathcal{D}$.

6 Problem 9 Toward bayesian inference

1. (4 points) Suppose you have a Bayes net over n nodes $\big(X_1,\ldots,X_n\big)$ and all the variables except X_i are **observed**. Using the chain rule and Bayes rule find an efficient algorithm to compute the probability:

$$P(x_i|x_1,\ldots,x_n) \tag{23}$$

Lets parition the nodes of the bayesian network according to their relation to X_i.

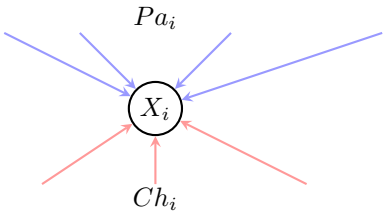
$$\begin{cases} \text{Pa}_i & \text{parent of } x_i \\ \text{Ch}_i & \text{childs of } x_i \\ \text{O}_i & \text{others} \end{cases} \tag{24}$$

Using the Bayes rule we obtain:

$$P(x_i|x_1,\ldots,x_n) = \frac{P\big(x_i|x_1\ldots,x_n\big)}{\sum_{x_i} P\big(x_i|x_1\ldots x_n\big)} \tag{25}$$

Using the nodes decomposition, we obtain

$$P(x_i|x_1,\ldots,x_n) = \frac{P(x_|Pa(x_i)\prod_{x\in Ch_i} P(x|Pa(x))}{\sum_{x_i} P(x_i|Pa(x_i)\prod_{x\in Ch_i} P(x|Pa(x)} \tag{26}$$



The entities should be computed for each values x_i in order to evaluate the partition normalizing factor.

2. (4 points) Find an efficient algorithm to generate random samples from the probability distribution defined in a Bayesian Network. You can assume access to a routine to draw a sample from *multinomial*. Let's consider the simple case $X \to Y$. By using the chain rule, $P(x,y) = P(x)P(y|x)$, we could draw a sample by using the procedure from a multinomial $\mathcal{M}$ as:

$$x = \mathcal{M}(X) \quad , \quad y = \mathcal{M}(Y|x)$$

Hence for any *BN* defined by the decomposition $P(x1,\ldots,x_n) = \prod P\big(x_i|Pa(x_i)\big)$, we start from the root[2] and for each x_i we draw a sample using

$$\mathcal{M}(x_i|Pa(x_i) = \text{sample}) \tag{27}$$

7 Programming Assignment

1. Since each pixel could take a values on $\{0,1\}$, the set of binary images of size 28×28 is $|I_{28,28}| = \mathbf{2}^{784}$
2. The number of parameter without conditioning is $2^{784}-1$.
3. For the Bayesian Net the probability distribution is written as:

$$P(X_1,\ldots,X_{784},Z_1,Z_2) = P(Z_1)P(Z_2)\prod_i P(X_i|Z_1,Z_2) \tag{28}$$

We need **784** conditional $P(X_i|Z_1,Z_2)$ probability for each X_i. The cardinality of a latent variable is 25. Therefore for each CPD, we need $25\times 25 - 1 = \mathbf{624}$ parameter.

Therefore, the number of parameters is $784\times 624 = 489216$.

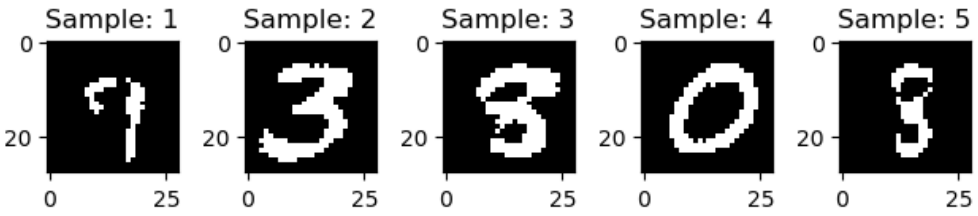


Figure 2: Sampling the Learned distribution

4. Here is the result, in Figure 2, of the sampling from the model:

5. The expected images are depicted on Figure 3.

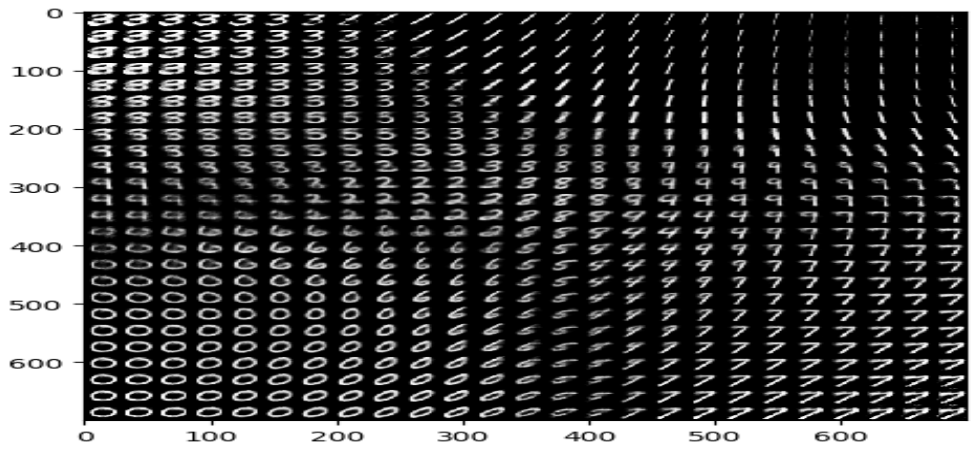


Figure 3: Excpected image for each latent value

6. We will classify the images according to the log-likehood

$$P(X_{1:784}) = \log \sum_{z_1} \sum_{z_2} p(z_1, z_2) \prod_i p(X_i|z_1, z_2) \tag{29}$$

   The proposed classifier will compute the mean and the standard deviation on the *validation* Dataset. For each image on the *Test*, An image is classified as normal is the lies within three standard deviation from the validation mean.

7. Question 7: visualize the scatter plot for the expected latent variables

[2] A DAG accepts a topological ordering

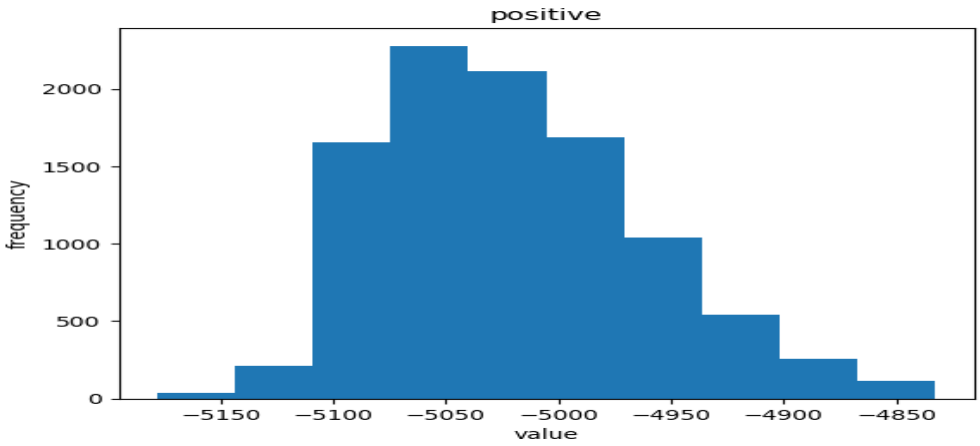


Figure 4: Histogram of the positive images

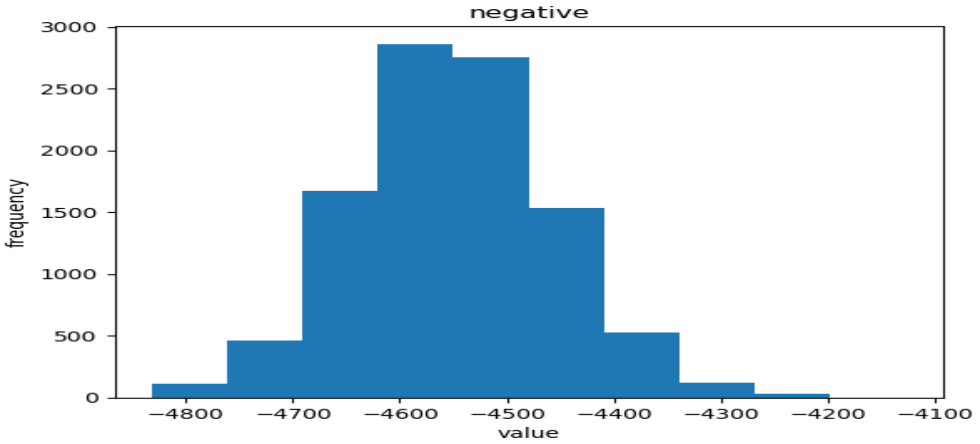


Figure 5: Histogram of the negative images

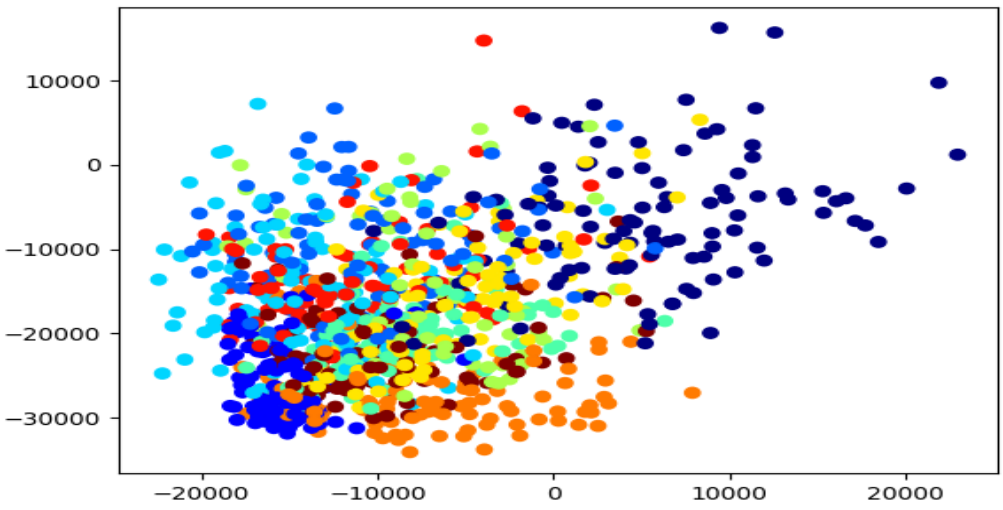


Figure 6: Distribution of the latent variables expectations