Question 1.

Consider modelling a spam filter by means of a joint probability distribution $P(Y, X_1, \ldots, X_n)$ such that

$$Y = \begin{cases} 1 \text{ if the message is a spam,} \\ 0 \text{ otherwise.} \end{cases} \qquad X_i = \begin{cases} 1 \text{ if the message contains the } i\text{-th English word,} \\ 0 \text{ otherwise.} \end{cases}$$

a) How many parameters do we need to store such distribution exactly? Do you have an estimate for n?

b) What problems may we encounter when trying to store the distribution exactly?

c) Do you have any ideas on how to improve the efficiency of the representation?

Answer:

a) The distribution is over $n+1$ binary random variables, hence there are 2^{n+1} parameters. Since the values must sum up to one, there are $2^{n+1}-1$ free parameters that we need to store.
A conservative estimate for the number of English words is $600,000$,[1]meaning that we will need $2^{600,001}-1$ parameters!
Just for reference, physicists estimate that there are $10^{80} \approx 2^{266}$ atoms in the known universe.[2]

b) Firstly, storage—obviously.
Secondly, computational instability—all numbers would basically be zero.

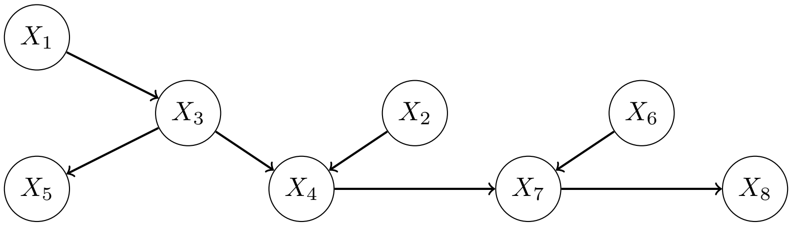
c) We can exploit relationships among some words.
Consider the words *car* and *drive*. Probabilities of those two words both being in a message are clearly correlated. However, the presence of the word *frog* likely won't give us any information about the presence of the word *pretzel*.
We can make use of *conditional independence* to model such relationships. That is exactly what **Bayesian networks** (and other probabilistic graphical models) do.
To get an idea about how much we can potentially save, imagine that presence of each word would be independent of all other words given that the message was spam or not. More formally,

$$P(y, x_1, \ldots, x_n) = P(y) \prod_{i=1}^{n} P(x_i|y).$$

Then, we would need one parameter to store $P(Y)$ and two parameters for each $P(X_i|Y)$. Overall, we would *only* need $2n+1$ parameters!
Clearly, the factorization above (*Naive Bayes*[3]) is an oversimplification, but it demonstrates that exponential savings are possible.

Question 2.

Consider the network (graph) below:



Decide the validity of the following statements:

[1] https://en.wikipedia.org/wiki/List_of_dictionaries_by_number_of_words (Oxford English Dictionary, 2nd edition)
[2] Helmenstine, Anne Marie. "How Many Atoms Exist in the Universe?" *ThoughCo., August 8, 2019.*
[3] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

a) $X_1 \perp\!\!\!\perp X_7 \mid X_3$

b) $X_1, X_5 \perp\!\!\!\perp X_6 \mid X_8$

c) $X_4 \perp\!\!\!\perp X_5 \mid X_1$

d) $X_1 \perp\!\!\!\perp X_2 \mid X_8$

e) $X_2 \perp\!\!\!\perp X_6$

f) $X_1 \perp\!\!\!\perp X_2, X_5$

Answer:

a) **True.** The only path X_1, X_3, X_4, X_7 is blocked by the observed X_3 (*causal chain*).

b) **False.** All paths go through X_7 (*common effect*) and they are active due to X_8 being observed.

c) **False.** The only path X_5, X_3, X_4 is active since X_3 is unobserved (*common cause*).

d) **False.** The path goes through X_4 (*common effect*) and it is active due to the descendant X_8 being observed.

e) **True.** The path X_6, X_7, X_4, X_2 is blocked by the unobserved X_7 (and its unobserved descendants).

f) **False.** The path X_5, X_3, X_1 is active since X_3 is unobserved.

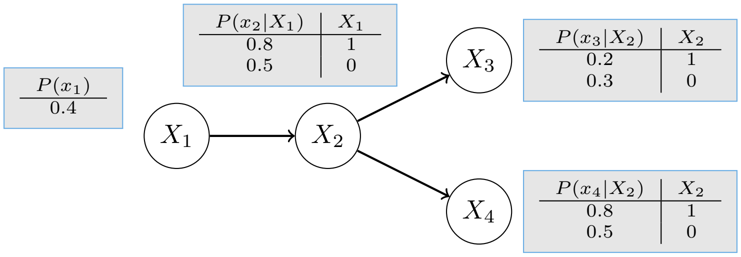
Question 3.

Consider the network below and compute

a) the marginal probability $P(X_3 = 0) = P(\neg x_3)$,

b) the pairwise marginal probability $P(X_2 = 1, X_3 = 0) = P(x_2, \neg x_3)$,

c) the conditional probability distribution $P(X_1 \mid X_2 = 1, X_3 = 0) = P(X_1 | x_2, \neg x_3)$.



$P(x_1)$
0.4

$P(x_2 \mid X_1)$	X_1
0.8	1
0.5	0

$P(x_3 \mid X_2)$	X_2
0.2	1
0.3	0

$P(x_4 \mid X_2)$	X_2
0.8	1
0.5	0

Answer:

Probability distribution induced by a BN is defined as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Par(X_i)).$$

To get rid of some variables, we need to "sum them out" (i.e., marginalize).

a)

$$\begin{aligned} P(X_3 = 0) &= \sum_{x_1=0}^{1} \sum_{x_2=0}^{1} \sum_{x_4=0}^{1} P(X_1 = x_1) \cdot P(X_2 = x_2 \mid X_1 = x_1) \cdot P(X_3 = 0 \mid X_2 = x_2) \cdot P(X_4 = x_4 \mid X_2 = x_2) \\ &= 0.6 \cdot 0.5 \cdot 0.7 \cdot 0.5 + 0.6 \cdot 0.5 \cdot 0.7 \cdot 0.5 + 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.2 + 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.8 \\ &+ 0.4 \cdot 0.2 \cdot 0.7 \cdot 0.5 + 0.4 \cdot 0.2 \cdot 0.7 \cdot 0.5 + 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.2 + 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.8 \\ &= 0.762 \end{aligned}$$

b)

$$\begin{aligned}P(X_2 = 1, X_3 = 0) &= \sum_{x_1=0}^{1}\sum_{x_4=0}^{1} P(X_1 = x_1) \cdot P(X_2 = 1 \mid X_1 = x_1) \cdot P(X_3 = 0 \mid X_2 = 1) \cdot P(X_4 = x_4 \mid X_2 = 1)\\ &= 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.2 + 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.8 + 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.2 + 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.8\\ &= 0.496\end{aligned}$$

c)

$$\begin{aligned}P(X_1 \mid X_2 = 1, X_3 = 0) &= \frac{P(X_1, X_2 = 1, X_3 = 0)}{P(X_2 = 1, X_3 = 0)}\\ P(X_1 = 1, X_2 = 1, X_3 = 0) &= \sum_{x_4=0}^{1} P(X_1 = 1) \cdot P(X_2 = 1 \mid X_1 = 1) \cdot P(X_3 = 0 \mid X_2 = 1) \cdot P(X_4 = x_4 \mid X_2 = 1)\\ &= 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.2 + 0.4 \cdot 0.8 \cdot 0.8 \cdot 0.8\\ &= 0.256\\ P(X_1 = 0, X_2 = 1, X_3 = 0) &= \sum_{x_4=0}^{1} P(X_1 = 0) \cdot P(X_2 = 1 \mid X_1 = 0) \cdot P(X_3 = 0 \mid X_2 = 1) \cdot P(X_4 = x_4 \mid X_2 = 1)\\ &= 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.8 + 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.2\\ &= 0.24\\ P(X_1 = 1 \mid X_2 = 1, X_3 = 0) &= \frac{0.256}{0.496} \approx 0.516\\ P(X_1 = 0 \mid X_2 = 1, X_3 = 0) &= \frac{0.24}{0.496} = 1 - \frac{0.256}{0.496} \approx 0.484\end{aligned}$$

Question 4.

Construct a Bayesian network (without CPTs) based on the following paragraph:

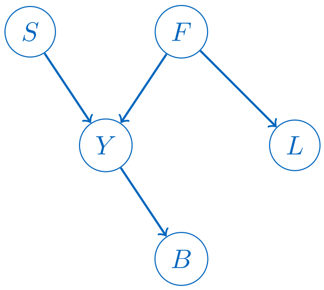
> When a family leaves their house, they often turn on the outdoor light. However, they also turn on the light when they are expecting a guest. The family has a dog, and they put it in the backyard when no one is home. They also put the dog there if it has bowel troubles. If the dog is in the backyard, it can probably be heard barking, although that could also be other dogs.

How many parameters do we need to represent such network?

How many parameters did we save compared to modelling the full joint distribution directly?

Answer:

The network **could** look like this:



where we have the following events (binary RVs):

- family in the house (F)
- outdoor light on (L)
- sick dog (S)
- dog in the backyard (Y)
- dog barking (B)

The network has $1(S) + 1(F) + 2(L) + 4(Y) + 2(B) = 10$ parameters.

Modelling the full joint directly would require $2^5 - 1 = 31$ parameters.

Thus, when using the BN above to represent the joint distribution, we only need *one third* of the originally required parameters!

Note:

The network above is also called *causal Bayes (belief) net*. The edges go from *causes* to their respective *effects*. However, the graph structure could be very different.

For one, we could add more edges between nodes (while not introducing a cycle!), reducing the number of assumptions about conditional independencies present in the joint probability distribution. For example, instead of storing just $P(B|Y)$, we could store $P(B|Y, L, F)$, effectively adding the edges (L, B) and (F, B) to the graph.

Second, we could change the edge orientation. For instance, we are storing $P(S), P(T)$ and $P(Y|S, T)$. However, we can also compute $P(Y|S)$ and $P(F|Y)$ and store those CPTs instead, reversing the edge (F, Y) to (Y, F).