Week 7: Data Analysis

Chapter 10: Bayesian Inference

- Degree of belief ↔ subjective probability
- Belief based on information ↔ conditional probability

$$\mathbb{P}[\text{event} \mid \text{information}]$$

- new information arrives:
  $\mathbb{P}[\text{event} \mid \text{original information} + \text{new information}]$
- When there are many events/information to be analyzed, how to perform the analysis systematically?
  - The basic tool is Bayesian network (also called causal network or Bayesian belief network).

Recaps of probability Theory

- Conditional Probability
- Bayes Formula

Example①: Conditional Probability

An integer between 1 and 12 is generated uniformly at random. Let X be the random variable that denotes the integer.

j	1	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}[X=j]$	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12
$\mathbb{P}[X=j \mid X \text{ is even}]$	0	1/6	0	1/6	0	1/6	0	1/6	0	1/6	0	1/6

- New information available: X is an even number
- So, X cannot be 1, 3, 5, 7, 9, 11.
- Formally, conditioned on X is even, the probability of X being one of those values is 0, i.e. for $j=1,3,5,7,9,11$. $\mathbb{P}[X=j \mid X \text{ is even}] = 0$
- For the rest possibilities, the probabilities raise in proportion.

- New information available: X is an even number.
- So X cannot be 1, 3, 5, 7, 9, 11.
- Another new information available: X is not divisible by 3.
- So X cannot be 6, 12.

j	2	4	8	10
P[X=j]	1/12	1/12	1/12	1/12
P[X=j | X is even & X is not divisible by 3]	1/4	1/4	1/4	1/4

Example ②: Conditional Probability

An integer between 1 and 12 is generated randomly following the distribution given in the table below. What is the conditional probability distribution if we know that X is even and not a multiple of 3?

j	1	2	3	4	5	6	7	8	9	10	11	12
P[X=j]	0.12	0.04	0.03	0.08	0.01	0.09	0.15	0.12	0.07	0.16	0.06	0.07
P[X=j | X is even & X is not divisible by 3]	0	0.1	0	0.2	0	0	0.3	0	0.4	0	0	

- After eliminating the impossible choices,
  - we raise the remaining probabilities in proportion.

For instance,

$$P[X=2 \mid X \text{ is even } \& X \text{ is not divisible by } 3] = \frac{0.04}{0.04+0.08+0.12+0.16} = 0.1$$

- What is P[X is divisible by 4 | X is even & X is not divisible by 3]?
  - Among the remaining possibilities, 4 and 8 are divisible by 4.
  - The desired conditional probability is $0.2+0.3 = 0.5$

Quiz① A fair die is thrown. Let X be its value. What is the conditional probability?

$P[X \text{ is not a multiple of } 3 \mid X \neq 5]$?

1	2	3	4	~~5~~	6
1/6	1/6	1/6	1/6	~~1/6~~	1/6
1/5	1/5	1/5	1/5		1/5

Therefore; $\frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{1+1+1}{5} = \frac{3}{5}$

Quiz② Two fair dice are thrown. Let X_1, X_2 be their values. What is the conditional probability?

$P[X_1 + X_2 \geq 8 \mid X_1 \geq 5]$?

X_1	1	2	3	4	5	6	* $X_1 \geq 5$
X_2	1	2	3	4	5	6	

* $X_1 + X_2 \geq 8$ ∴ $(X_1, X_2) = (5,3), (5,4), (5,5), (5,6)$
$(6,2), (6,3), (6,4), (6,5), (6,6)$

∴ Ans $= \frac{9}{12} = \frac{3}{4}$

Bayes Formula

- The probability of event A conditioned on event E is denote by

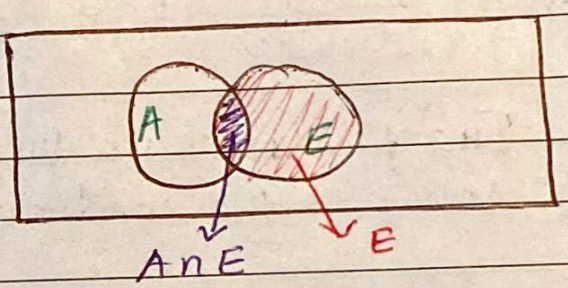
$$P[A|E]$$

- The Bayes Formula:

$$P[A|E] = \frac{P[A \text{ and } E]}{P[E]} = \frac{P[A \cap E]}{P[E]}$$

Provided that $P[E] > 0$. or if $P[E] \geq 0$

- Venn diagram:



- We can rewrite it as $P[A|E].P[E] = P[A \text{ and } E]$
- Since the formula holds for any A, E,
  - we can swap A, E to get

$$P[E|A].P[A] = P[E \cap A]$$

Thus, we have

$$P[A|E].P[E] = P[E|A].P[A]$$

$$P[A|E] = P[E|A] \cdot \frac{P[A]}{P[E]}$$

Example: Let X be a random email.
Let A denote the event X is a spam, and E denote the event X contains the word "free".
As an admin, you are wondering if it is good idea to filter out all emails containing the word "free".

Thus, you are interested to know $P[A|E]$.
If this probability is very close to 1, then the filter is justified.
You do not have data to directly estimate $P[A|E]$, but fortunately, other surveys have estimated $P[E|A]$, $P[A]$ and $P[E]$, say
$P[E|A] = 0.2$, $P[A] = 0.8$ and $P[E] = 0.1608$.
Then

$$P[A|E] = P[E|A] \cdot \frac{P[A]}{P[E]} = 0.2 \cdot \frac{0.8}{0.1608} = 0.995025.$$

- The Bayes formula: If $P[E] > 0$,

$$P[A|E] = \frac{P[A \cap E]}{P[E]}$$

- Suppose that $A_1, A_2, \ldots, A_m$ form a partition of the whole sample space.
  - Then we can write E as disjoint union of
    $E \cap A_1, E \cap A_2, \ldots, E \cap A_m$
  - Thus, $P[E] = \sum_{i=1}^{m} P[E \cap A_i]$
  - Then by the Bayes formula, $P[E \cap A_i] = P[E|A_i] \cdot P[A_i]$
  - Hence, we have the following formula:

$$P[A_i|E] = \frac{P[A_i \cap E]}{P[E]} = \frac{P[E|A_i] \cdot P[A_i]}{\sum_{j=1}^{m} P[E|A_j] \cdot P[A_j]}$$

Bayes Formula

$$P[A_i|E] = \frac{P[E|A_i] \cdot P[A_i]}{\sum_{j=1}^{m} P[E|A_j] \cdot P[A_j]}$$

Example: Let X be a random email.

- Let A denote the event X is a spam,
- and E denote the event X contains the word "free".
- As an admin, you are wondering if it is good idea to filter out all emails containing the word "free".

- It is known that 80% of all emails are spam, i.e. $P[A] = 0.8$
- It is also known that among emails which are spam, 20% of them contain the word "free", i.e. $P[E|A] = 0.2$.
- Yet, among emails which are not spam, only 1% of them contain the word "free", i.e. $P[E|\text{not } A] = 0.01$.

By the above formula, we have

$$P[A|E] = \frac{P[E|A] \cdot P[A]}{P[E|A] \cdot P[A] + P[E|\text{not}A] \cdot P[\text{not}A]}$$

$$= \frac{0.2 \times 0.8}{0.2 \times 0.8 + 0.01 \times 0.2} = 0.987654.$$

- Filtering out all emails containing the word "free" does not seem like a good idea.
- The admin might want to design a more sophisticated filter rule.

Summary of Bayes Formula

- The Bayes formula: if $P[E] > 0$,

$$P[A|E] = \frac{P[A \cap E]}{P[E]}$$

- Other useful formulae can be derived from this, including
  - $P[A|E] \cdot P[E] = P[E|A] \cdot P[A]$
  - if $A_1, A_2, \ldots, A_m$ form a partition of the whole sample space, then

$$P[A_i|E] = \frac{P[E|A_i] \cdot P[A_i]}{\sum_{j=1}^{m} P[E|A_j] \cdot P[A_j]}$$

Bayesian Network

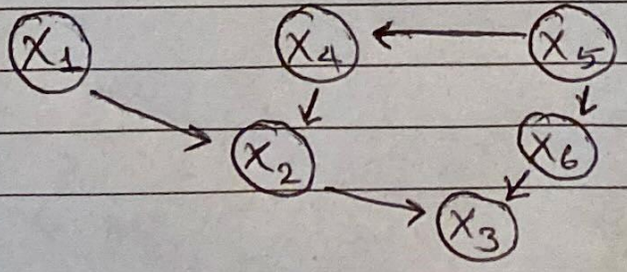
- A graph-theoretic definition.
- Number of parameters for specification.

- We covered about conditional probability & Bayes formula.
  - But Why covering them?
- Recall that: we want to use data to gauge how likely an event will occur when given some information (evidence).

- There can be many relevant information, and we need a systematic way to organize them and use them for gauging how likely an event will occur.

- A short answer is: Bayesian network
  - is the main tool to organize, the Bayes formula is the main engine for using the network.

- In a Bayesian network,
  - There are a number of nodes (vertices).
  - Each node represents a random variable, which we usually denote by X_i.
  - To keep the discussion simple, we assume each random variable has value 0 or 1.
  - There are some directed edges between the nodes.
  - Requirement: there is no directed cycle in a Bayesian network.
- No directed cycle: you cannot start from one node, follow the arrows, and go back to itself.



- The following means:
  - the randomness of X_j directly depends on the value of X_i.
  - (Read: $X_i \rightarrow X_j$ as "X_i influences X_j".)

Specification

When $X_i = 0$:

$$P[X_j = 0 \mid X_i = 0] = a$$
$$P[X_j = 1 \mid X_i = 0] = 1 - a.$$

When $X_i = 1$:

$$P[X_j = 0 \mid X_i = 1] = b$$
$$P[X_j = 1 \mid X_i = 1] = 1 - b$$

We need 2 parameters (a & b) to represent $P[X_j \mid X_i]$.

- The following means:
  - The value of X_k directly depends on the values of X_i, X_j.

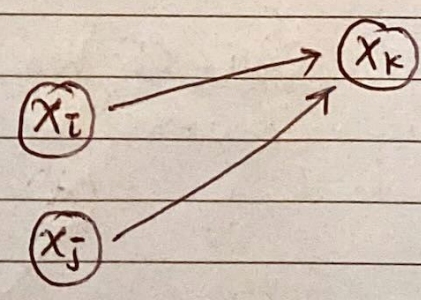
Specification

$$P[X_k = 0 \mid X_i = 0, X_j = 0] = a$$
$$P[X_k = 1 \mid X_i = 0, X_j = 0] = 1 - a$$

$$P[X_k = 0 \mid X_i = 0, X_j = 1] = b$$
$$P[X_k = 1 \mid X_i = 0, X_j = 1] = 1 - b$$

$$P[X_k = 0 \mid X_i = 1, X_j = 0] = c$$
$$P[X_k = 1 \mid X_i = 1, X_j = 0] = 1 - c$$

$$P[X_k = 0 \mid X_i = 1, X_j = 1] = d$$
$$P[X_k = 1 \mid X_i = 1, X_j = 1] = 1 - d$$



We need 4 parameters (a, b, c & d) to represent $P[X_k \mid X_i, X_j]$.

Example:

X_5: education level of a person (high/low)

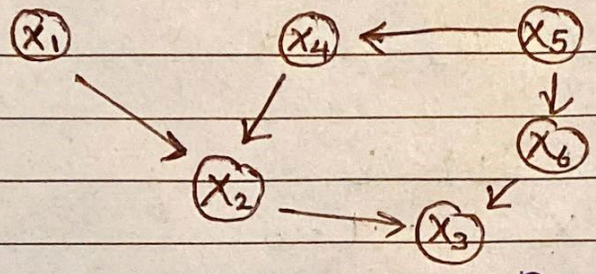
X_4: the person trusts vaccine

X_1: the person sees government's campaign

X_2: the person takes a vaccine

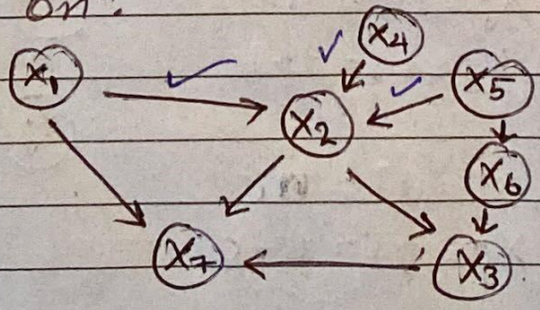
X_6: the person wears mask outdoors

X_3: the person catches COVID



- You may argue that this Bayesian Network is not correct/perfect,
  e.g. X_1 should have a directed edge to X_4 instead of X_2, or the gender and age of the person should be taken into account.
- This belongs to a bigger phenonmenom of stat and science:
  - All models are wrong, but some are useful.

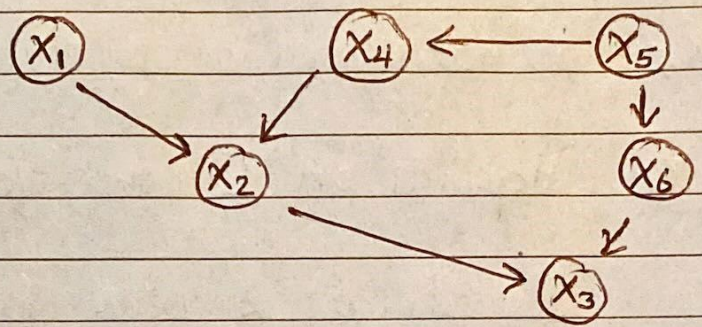
Quiz ③ Which random variables does X_2 directly depends on?



(A) X_1, X_4, X_5

- In general, when a random variable directly depends on q other random variables, then
  - we need 2^q parameters to specify the conditional dependence.
  - (This is true only for binary random variables.).

Example: In this Bayesian network,
- X_1, X_5 each depends on no other random variable;
- X_4, X_6 each depends on one other random variable;
- X_2, X_3 each depends on two other random variables.



- Thus, the total number of parameters needed for specifying the whole Bayesian network is

$$2^0 \times 2 + 2^1 * 2 + 2^2 \times 2 = 14$$

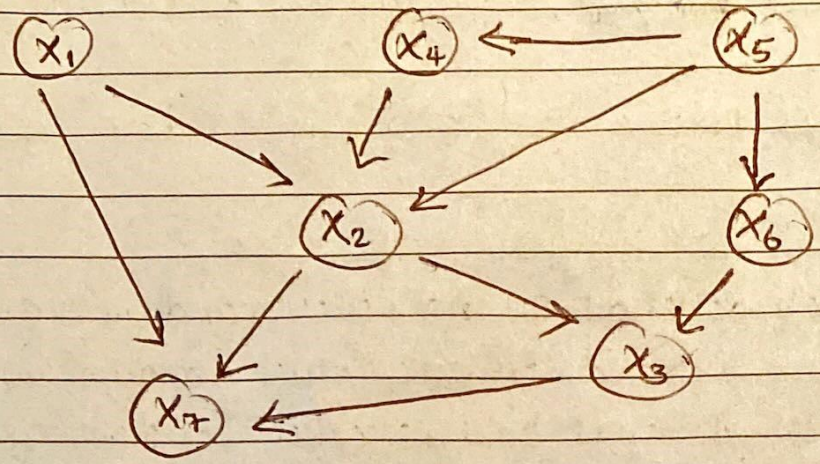
- In Data Analysis, these parameters are estimated via techniques you learnt in Density Estimation, e.g. histogram.

- In contrast, in the most general model, we need to have an estimate for each possibility of $(X_1, X_2, X_3, X_4, X_5, X_6)$, e.g.

$$P[X_1=0, X_2=1, X_3=0, X_4=1, X_5=1, X_6=1] = ???$$

- There are $2^6 = 64$ possibilities, and hence 63 parameters are needed.
  (It is 64−1 because the probabilities need to sum up to 1.)
- Be reminded that when a model has more parameters, it is more flexible (more degrees of freedom). If there are too many parameters, overfitting might occur.

Quiz: How many parameters are needed to specify the Bayesian network?



$2^0 \times 2$ X_1, X_5 do not depend on any other variable

$2^1 \times 2$ X_4, X_6 depends on 1 other variable

$2^2 \times 1$ X_3 depends on 2 other variable

$2^3 \times 2$ X_2, X_7 depends on 3 other variable

Therefore;

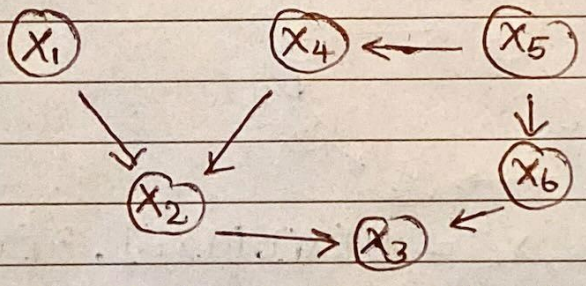
$2^0 \times 2 + 2^1 \times 2 + 2^2 * 1 + 2^3 \times 2 = 2 + 4 + 4 + 16 = 26$

Bayesian Inference

- Now you should have a good idea what Bayesian network is.
  - But how to use it?
- Our target is to perform Bayesian inference, i.e. compute

$$P[X_{i_1}=x_{i_1}, \dots, X_{i_k}=x_{i_k} \mid X_{j_1}=x_{j_1}, \dots, X_{j_l}=x_{j_l}]$$

Example: Recall that
- X_5: education level of a person (high/low)
- X_4: the person trusts vaccine
- X_1: the person sees government's campaign
- X_2: the person takes a vaccine
- X_6: the person wears mask outdoors
- X_3: the person catches COVID



- We want to find out the effectiveness of government's campaign in reducing COVID cases among non-highly educated people.
- Thus we are interested in comparing
  $P[X_3=0 \mid X_1=1, X_5=0]$ with
  $P[X_3=0 \mid X_1=0, X_5=0]$.

- Our target is to perform Bayesian inference, i.e. compute

$$P[X_{i_1}=x_{i_1}, \dots, X_{i_k}=x_{i_k} \mid X_{j_1}=x_{j_1}, \dots, X_{j_l}=x_{j_l}].$$

- For graph which are polytrees,
- For general graphs, there exists more involved algorithms that perform Bayesian inference exactly, e.g. Junction tree algorithm.

Configuration problem

- We will start with the configuration problem
  i.e. Computing
  $$P[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n].$$

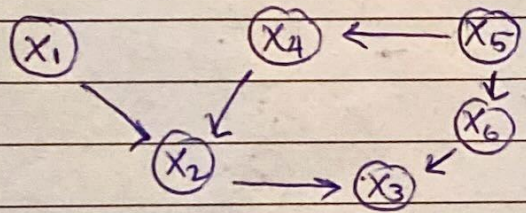
Note that: all random variables are involved.

Example:

Recall that:

- A Bayesian network must have no directed cycle,
  i.e. there is a causality ordering of the random variables.

- The Bayesian network can be represented algebraically by sorting out the causality ordering.



$$P[X_1, X_2, X_3, X_4, X_5, X_6] = P[X_1] \cdot P[X_5] \cdot P[X_4|X_5] \cdot P[X_6|X_5] \cdot P[X_2|X_1, X_4] \cdot P[X_3|X_2, X_6]$$

- with this, we can solve the atomic problem easily,
  e.g.

$$P[X_1=0, X_2=1, X_3=0, X_4=1, X_5=0, X_6=1] = P[X_1=0] \cdot P[X_5=0] \cdot P[X_4=1|X_5=0] \cdot P[X_6=1|X_5=0] \cdot P[X_2=1|X_1=0, X_4=1] \cdot P[X_3=0|X_2=1, X_6=1]$$

Bayesian Inference

- We have learnt how to solve one configuration problem.
- This is the building block for solving other problems.

Example:

- Suppose a Bayesian network has 6 random variables, $X_1, X_2, \ldots, X_6$.

- To compute

$$P[X_1=0, X_2=0, X_4=1, X_6=1]$$

- We list out all configurations that satisfy the conditions as below:

$$\begin{aligned}P[X_1=0, X_2=0, X_4=1, X_6=1] &= P[X_1=0, X_2=0, X_3=0, X_4=1, X_5=0, X_6=1] \\ &+ P[X_1=0, X_2=0, X_3=0, X_4=1, X_5=1, X_6=1] \\ &+ P[X_1=0, X_2=0, X_3=1, X_4=1, X_5=0, X_6=1] \\ &+ P[X_1=0, X_2=0, X_3=1, X_4=1, X_5=1, X_6=1]\end{aligned}$$

- Computing each probability in the RHS
  - is a configuration problem.
  - In other words, computing
    $P[X_1=0, X_2=0, X_4=1, X_6=1]$
    can be reduced to solving 4 configuration problems.

- Recall that: Our target is to perform Bayesian inference.
  i.e. Compute
$$P[X_{i_1}=x_{i_1}, \ldots, X_{i_k}=x_{i_k} \mid X_{j_1}=x_{j_1}, \ldots, X_{j_l}=x_{j_l}].$$

Example: By the Bayes formula. $P(A|E) = \frac{P(A \cap E)}{P(E)}$

$$P[X_3=0 \mid X_1=1, X_5=0] = \frac{P[X_1=1, X_3=0, X_5=0]}{P[X_1=1, X_5=0]}$$

- The numerator can be computed by solving $2^3=8$
- while the denominator can be computed by solving $2^4=16$ configuration problems.

- When the number of random variable increase,
  - the number of configuration problems we want to solve can increase exponentially.
  - This is why we said it is a slow but simple method.

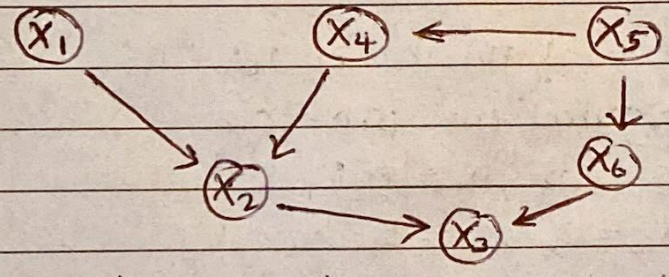
Quiz: Suppose a Bayesian network has 10 random variables $X_1, X_2, \ldots, X_{10}$. To compute

$$P[X_4 = 1, X_6 = 1, X_8 = 0, X_9 = 0],$$

How many configuration problems we will need to solve?

Example: Finally, the Bayes formula is coming into play...

- We will present an artificial example of Bayesian inference; the Bayesian network is given below.



- We want to find out the effectiveness of government's campaign on reducing COVID cases among non-highly educated people.
- Thus we are interested in comparing

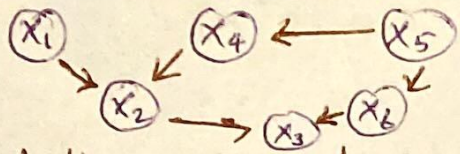
$$P[X_3 = 0 | X_1 = 1, X_5 = 0] \text{ with } P[X_3 = 0 | X_1 = 0, X_5 = 0]$$

Main Step 1.

Recall that: this network is specified by 14 parameters.

- These paramters are derived from data using techniques in density estimation.

  e.g. Histogram.



- We list the parameters we use in this example (these parameters are fabricated, not reflecting the truth)

$$P[X_{1\sim6}] = P[X_3|X_2,X_6]\cdot P[X_2|X_1,X_4]\cdot P[X_6|X_5]\cdot P[X_4|X_5]\cdot P[X_5]\cdot P[X_1]$$

$P[X_1=0]=0.3$
$P[X_5=0]=0.6$

$P[X_6=0|X_5=0]=0.9$
$P[X_6=0|X_5=1]=0.4$

$P[X_3=0|X_2=0,X_6=0]=0.6$
$P[X_3=0|X_2=0,X_6=1]=0.3$
$P[X_3=0|X_2=1,X_6=0]=0.1$
$P[X_3=0|X_2=1,X_6=1]=0.$

$P[X_4=0|X_5=0]=0.8$
$P[X_4=0|X_5=1]=0.2$

$P[X_2=0|X_1=0,X_4=0]=0$
$P[X_2=0|X_1=0,X_4=1]=0.8$
$P[X_2=0|X_1=1,X_4=0]=0.4$
$P[X_2=0|X_1=1,X_4=1]=0.9$

Main Step ②: Use the Bayes formula.

$$P[X_3=0|X_1=1,X_5=0] = \frac{P[X_1=1,X_3=0,X_5=0]}{P[X_1=1,X_5=0]}$$

Main step ③:
- For the numerator, write it as a sum of probabilities of configurations.
- Compute each of them using the red formula at the top.

$P[X_1=1,X_3=0,X_5=0] = P[X_1=1,X_2=0,X_3=0,X_4=0,X_5=0,X_6=0]$ 0.072576
$+P[X_1=1,X_2=0,X_3=0,X_4=0,X_5=0,X_6=1]$ 0.004032
$+P[X_1=1,X_2=0,X_3=0,X_4=1,X_5=0,X_6=0]$ 0.040824
$+P[X_1=1,X_2=0,X_3=0,X_4=1,X_5=0,X_6=1]$ 0.002268
$+P[X_1=1,X_2=1,X_3=0,X_4=0,X_5=0,X_6=0]$ 0.018144
$+P[X_1=1,X_2=1,X_3=0,X_4=0,X_5=0,X_6=1]$ 0
$+P[X_1=1,X_2=1,X_3=0,X_4=1,X_5=0,X_6=0]$ 0.000756
$+P[X_1=1,X_2=1,X_3=0,X_4=1,X_5=0,X_6=1]$ 0

- As on the previous slide, you can compute both
  - the numerator and
  - the denominator on RHS.
- so you should try to write a program to ask computer to do it for you. I did so!
  (see the file bayes_example.py)

- We have

$$P[X_3=0 \mid X_1=1, X_5=0] = \frac{P[X_1=1, X_3=0, X_5=0]}{P[X_1=1, X_5=0]} = \frac{0.1386}{0.42}$$

$$= 0.33\,4.$$

Analogously, you can also comput

$$P[X_3=0 \mid X_1=0, X_5=0] = \frac{P[X_1=0, X_3=0, X_5=0]}{P[X_1=0, X_5=0]} = \frac{0.030024}{0.18}$$

$$= 0.1668$$

Thus, we are interested in comparing

0.33 with 0.1668.