SI140 Discussion 03

Li Zeng, Tao Huang, Xinyi Liu

ShanghaiTech University, China
{zengli,huangtao1,liuxy10}@shanghaitech.edu.cn

1 Basis of Conditional Probability

We have introduced probability as a language for expressing our degrees of belief or uncertainties about events. Whenever we observe new evidence (i.e., obtain data), we acquire information that may affect our uncertainties. A new observation that is consistent with an existing belief could make us more sure of that belief, while a surprising observation could throw that belief into question. Conditional probability is the concept that addresses this fundamental question: how should we update our beliefs in light of the evidence we observe?
Due to the central importance of conditioning, both as the means by which we update beliefs to reflect evidence and as a problem-solving strategy, we say that

Conditioning is the soul of statistics and inferences

Definition 1 (Conditional probability). *If A and B are event with $P(B) > 0$, then the conditional probability of A given B, denoted by $P(A|B)$, is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- $P(A)$: Prior probability (old belief) of A .
- B: Something happens that your inference is based on (also called evidence or observation).
- $P(A|B)$: Posterior probability (updated belief) of A .
- $P(A|B)P(B) = P(B|A)P(A) = P(A \cap B)$

Example 1 (How Iqiyi Learns the Recommendation). Given that you watched this movie, what is the probability that you will watch another movie?

Suppose we are looking for the probability that a user watches the movie Life is Beautiful (call this event E). One possible approach to calculating $P(E)$ is from the frequentist definition of probability: Let n be the number of users on Netflix, and let $n(E)$ be the number of users who have watched this particular movie. since n is large (there are a lot of people on Netflix!) we can compute $P(E) \approx \frac{n(E)}{n}$ Now suppose we are looking for $P(E \mid F)$, the probability that a user watches Life is Beautiful given they watched another movie, Amelie. By the definition of conditional probability, $P(E \mid F) = \frac{P(EF)}{P(F)}$, and by the frequentist definition of probability $P(F) \approx \frac{n(F)}{n}$ and $P(EF) \approx \frac{n(EF)}{n}$, where $n(F)$ and $n(EF)$ are the numbers of users who have watched Life is Beautiful and both movies, respectively. Therefore $P(E \mid F) \approx \frac{n(EF)}{n(F)}$ The purpose of this example is to show that this large denominator of people on Netflix will go away in conditional probability, and furthermore that the statistics of $P(E)$ and $P(E \mid F)$ can be different.

Theorem 1 (Chain Rule). *For any events $A_1, ..., A_n$ with positive probabilities,*

$$P(A_1, ..., A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, ..., A_{n-1})$$

Theorem 2 (The Law of Total Probability). *Let $A_1, ..., A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then*

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Exercise 1. Your friend Ethan wants to find the biggest flower in the field. In the field, there are n flowers. The size of all the flowers can be ranked unambiguously if all are seen. Ethan will walk across this field, with each order of flowers being equally likely. Ethan decides not to walk one meter back because he is too tired. He also only gives him one chance to pick the flower. He wants to have the highest probability of selecting the actual biggest flower.

(a) Ethan will check out $r-1$ flowers without touching them so that he can have a good calibration. M is the biggest flower in the $r-1$ flowers. He then selects the first subsequent flowers that is better than M (and the last flower if nothing was picked till then). For an arbitrary cutoff r, derive the probability that the biggest flower is actually selected.
Hint: Use LOPT to solve the problem.

(b) Now compute the probability when n goes to ∞, assume $r \sim \mathcal{O}(n)$.
Hint: you might need to do a bit of calculus here. For example, when $n \to \infty, \frac{1}{n}$ can be dt in an integral. Here t is a new variable we introduce.

(c) Take the derivative with respect to your expression from (b) and tell Ethan when should he stop calibration. (what is the optimal r*?)

Solution 1. (a) Since the order of flowers is random, so

$$P(\text{ flower i is the best}) = \frac{1}{n} \quad \forall i \in [1, 2, \cdots, n]$$

$$\forall i \in \{1, 2, \cdots, r-1\}, P(\text{flower i is selected}) = 0$$
$$\Rightarrow P(\text{ flower } i \text{ is selected } \mid \text{i is the best}) = 0$$

$$\forall i \in \{r, \cdots, n\}, P(\text{ flower } i \text{ is selected } \mid i \text{ is the best})$$
$$= P(\text{the second biggest flower is in the first r-1 ones}) = \frac{r-1}{i-1}$$

Using LOTP, we have

$$\begin{aligned} P(r) &= \sum_{i=1}^{n} P(\text{ flower } i \text{ is selected and it is the biggest}) \\ &= \sum_{i=1}^{n} P(\text{ flower } i \text{ is selected } \mid i \text{ is the best})\ P(\text{ flower } i \text{ is the best}) \\ &= \sum_{i=r}^{n} P(\text{ flower } i \text{ is selected } \mid i \text{ is the best}) P(\text{ flower } i \text{ is the best }) \\ &= \frac{1}{n} \sum_{i=r}^{n} \frac{r-1}{i-1} \end{aligned}$$

(b)

$$\begin{aligned} P(r) &= \frac{1}{n} \sum_{i=r}^{n} \frac{r-1}{i-1} \\ &= \left(\frac{r-1}{n}\right) \sum_{i=r}^{n} \frac{n}{i-1} \frac{1}{n} \end{aligned}$$

assume $r \sim \mathcal{O}(n)$, let $\frac{r}{n} = p$,

$$\begin{aligned} P(r) &= \left(p - \frac{1}{n}\right) \sum_{i=1}^{n} \frac{n}{i-1} \frac{1}{n} \\ &= p \int_{p}^{1} \frac{1}{t} dt \quad (n \to \infty) \\ &= P(\ln 1 - \ln p) = \frac{r}{n} \ln\left(\frac{n}{r}\right) \end{aligned}$$

(c) We take derivative of $P(r)$ w,r,t r,

$$\begin{aligned}\frac{d(P(r))}{dr} &= \frac{d}{dr}\left(\frac{1}{n}(\ln n \cdot r - r\ln r)\right)\\ &= \frac{\ln n - \ln r - 1}{n}\\ &= 0\\ \Rightarrow \quad r^* &= \frac{n}{e}\end{aligned}$$

Lemma 1 (Bayes + LOTP = Inferences).

$$P(A_i|B) = \frac{P(A_i,B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Note: You may find that the posterior belief of A_i is a weighted ratio of all possible A_is, with the weight of a conditional probability – the probability of the evidence B given A_i.

Exercise 2 (A Girl Named Lulu).

(a) A family has two children. Given that one of them is a girl, what is the probability that both are girls?
(b) A family has two children. Given that one of them is a girl named Lulu, what is the probability that both are girls?

Hint: The key of this question is that Lulu is an unusual name for Girl.

Solution 2. (a) Generally, there are four possible gender combinations for the two children, which are all equally likely: $(\mathrm{B},\mathrm{B}), (\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})$
Knowing that one of the children is a girl, only three options remain: $(\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})$. Therefore, the probability of (G,G) is $\frac{1}{3}$.
(b) An frequentist way to calculate this: assume 1 in 1000 girls is named Lulu. Of 100,000 families with two children, 75,000 will have at least one girl: 50,000 will have a girl and a boy, and 25,000 will have two girls. Of the 50,000 girl/boy families, we expect 50 to have a girl named Lulu. Of the 25,000 girl/girl families, we expect 50 to have a girl named Lulu: 25 where the first-born is Lulu, 25 where the second born is Lulu. The probability is thus $\frac{1}{2}$.

Exercise 3. There are n urns of which the r-th contains $r-1$ red balls and $n-r$ magenta balls. You pick an urn at random and remove two balls at random without replacement. Find the probability that:

1. the second ball is magenta,
2. the second ball is magenta, given that the first is magenta.

Solution 3. Let C_i be the color of the i-th ball picked, and use the obvious notation.

1. Since each urn contains the same number $n-1$ of balls, the second picked ball picked is equally likely to be any of the $n(n-1)$ available. One half of these balls are magenta, whence $P(C_2 = M) = 0.5$
2. By conditioning on the choice of urn,

$$P(C_2 = M|C_1 = M) = \frac{P(C_1, C_2 = M)}{P(C_1 = M)} = \sum_{r=1}^{n} \frac{\frac{(n-r)(n-r-1)}{n(n-1)(n-2)}}{\frac{1}{2}} = \frac{2}{3}$$

2 Conditional Probabilities are Probabilities

- Conditional probabilities are between 0 and 1.
- $P(S|E) = 1, P(\phi|E) = 0$.
- If $A_1, A_2, ...$ are disjoint, then $P(\cup_{j=1}^{\infty} P(A_j|E))$.
- $P(A^c|E = 1 - P(A|E))$.
- Inclusion-exclusion: $P(A \cup B|E) = P(A|E) + P(B|E) - P(A \cap B|E)$.

Theorem 3 (Bayes' Rule with Extra Conditioning). *Provided that* $P(A \cap E) > 0$ *and* $P(B \cap E) > 0$, *we have*

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

Theorem 4 (LOTP with Extra Conditioning). *Let* $A_1, ..., A_n$ *be a partition of the sample space* S *(i.e., the* A_i *are disjoint events and their union is* S*), with* $P(A_i \cap E) > 0$ *for all* i. *Then*

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i, E)P(A_i|E)$$

Exercise 4. You are lost in the National Park of **Bandrika**[1]. Tourists comprise two-thirds of the visitors to the park, and give a correct answer to requests for directions with probability $\frac{3}{4}$. (Answers to repeated questions are independent, even if the question and the person are the same.) If you ask a Bandrikan for directions, the answer is always false.

(a) You ask a passer-by whether the exit from the Park is East or West. The answer is East. What is the probability this is correct?
(b) You ask the same person again, and receive the same reply. Show the probability that it is correct is $\frac{1}{2}$
(c) You ask the same person again, and receive the same reply. What is the probability this is correct?
(d) You ask the fourth time, and receive the answer East. Show that the probability it is correct is $\frac{27}{70}$.
(e) Show that, had the fourth answer been West instead , the probability that East is nevertheless correct is $\frac{9}{10}$

Solution 4. Let S_r denote the event that you receive r similar answers, and T the event that they are correct. Denote the event that your interlocutor is a tourist by V. Then $T \cap V^c = \emptyset$, and

$$P(T|S_r) = \frac{P(T \cap V \cap S_r)}{P(S_r)} = \frac{P(T \cap S_r|V)P(V)}{P(S_r)}$$

Hence:

1. $P(T|S_1) = \frac{3}{4} \times \frac{2}{3}/1 = \frac{1}{2}$
2. $P(T|S_2) = \dfrac{(\frac{3}{4})^2 \times \frac{2}{3}}{\frac{1}{3} + \frac{2}{3} \times \left[(\frac{3}{4})^2 + (\frac{1}{4})^2\right]} = \frac{1}{2}$
3. $P(T|S_3) = \dfrac{(\frac{3}{4})^3 \times \frac{2}{3}}{\frac{1}{3} + \frac{2}{3} \times \left[(\frac{3}{4})^3 + (\frac{1}{4})^3\right]} = \frac{9}{20}$
4. $P(T|S_4) = \dfrac{(\frac{3}{4})^4 \times \frac{2}{3}}{\frac{1}{3} + \frac{2}{3} \times \left[(\frac{3}{4})^4 + (\frac{1}{4})^4\right]} = \frac{27}{70}$

[1] A fictional country made famous in the Hitchcock film 'The Lady Vanishes'.

5. If the last answer differs, then the speaker is surely a tourist, so the required probability is

$$\frac{\frac{1}{4}\times(\frac{3}{4})^3}{\frac{1}{4}\times(\frac{3}{4})^3+\frac{3}{4}\times(\frac{1}{4})^3}=\frac{9}{10}$$

Exercise 5. **Mr. Bayes goes to Bandrika**. Tom is in the same position as you were in the previous problem, but he has reasons to believe that, with probability ϵ, East is the correct answer. Show that:

(a) what ever answer is first received, Tom continues to believe that East is correct with probability ϵ,
(b) if the first two replies are the same (that is, either WW or EE), Tom continues to believe that East is correct with probability ϵ,
(c) after three like answers, Tom will calculate as follows, in the obvious notation:

$$P(\text{East correct}|\text{EEE})=\frac{9\epsilon}{11-2\epsilon},\quad P(\text{East correct}|\text{WWW})=\frac{11\epsilon}{11+2\epsilon}.$$

Solution 5. Let E (respectively W) denote the event that the answer East (respectively West) is given.

1. Using the conditional probability,

$$P(\text{East correct}|E)=\frac{\epsilon P(E|\text{East correct})}{P(E)}=\frac{\epsilon\cdot\frac{2}{3}\cdot\frac{3}{4}}{\frac{1}{2}\epsilon+(\frac{2}{3}\cdot\frac{1}{4}+\frac{1}{3})(1+\epsilon)}=\epsilon$$

$$P(\text{East correct}|W)=\frac{\epsilon(\frac{2}{3}\cdot\frac{1}{4}+\frac{1}{3})}{\epsilon(\frac{1}{6}+\frac{1}{3})+\frac{2}{3}\cdot\frac{3}{4}(1-\epsilon)}=\epsilon$$

2. Likewise, one obtains for the answer EE,

$$\frac{\epsilon\cdot\frac{2}{3}(\frac{3}{4})^2}{\epsilon\cdot\frac{2}{3}(\frac{3}{4})^2+(1-\epsilon)\left(\frac{2}{3}(\frac{1}{4})^2+\frac{1}{3}\right)}=\epsilon$$

and for the answer WW,

$$\frac{\epsilon\left(\frac{2}{3}(\frac{1}{4})^2+\frac{1}{3}\right)}{\epsilon\cdot\frac{3}{8}+(1-\epsilon)\cdot\frac{3}{8}}=\epsilon$$

3. Similarly for EEE,

$$\epsilon\frac{2}{3}(\frac{3}{4})^2\left[\epsilon(\frac{2}{3})(\frac{3}{4})^3+(1-\epsilon)\left((\frac{2}{3})\cdot(\frac{1}{4})^3+\frac{1}{3}\right)\right]=\frac{9\epsilon}{11-2\epsilon}$$

3 Independence of Events

3.1 Definition of Independence

Definition 2 (Independence of two events). *Events A and B are independent if and only if*

$$P(A\cap B)=P(A)P(B).$$

If $P(A)>0$ and $P(B)>0$, then this is equivalent to

$$P(A|B)=P(A)\quad or\quad P(B|A)=P(B)$$

Definition 3 (Independence of multiple events). *In general, n events $E_1,E_2,\ldots,E_n$ are independent if for every subset with r elements (where $r\leq n$) it holds that:*

$$P\left(E_{i_1},E_{i_2},\ldots,E_{i_r}\right)=P\left(E_{i_1}\right)P\left(E_{i_2}\right)\ldots P\left(E_{i_r}\right)$$

3.2 Conditional Independence

Definition 4 (Conditional Independence). *Events A and B are said to be conditionally independent given E if*

$$P(A \cap B|E) = P(A|E)P(B|E)$$

3.3 Relation between Independence & Conditional Independence

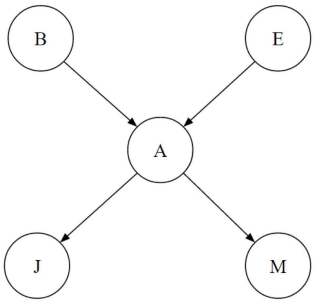
- Conditional Independent does not imply Independent.
- Independent does not imply Conditional Independent.

Example 2 (Bayes Nets). We formally define a Bayes Net as consisting of:

- A directed acyclic graph of nodes, one per variable X.
- A conditional distribution for each node $P(X \mid A_1 \ldots A_n)$, where A_i is the i^{th} parent of X, stored as a conditional probability table or CPT.
- **Fundamental assumption of BN structure:** Each node is **conditionally independent** of all its ancestor nodes in the graph, given all of its parents.

Exercise 6. Bayes nets sometimes is a efficient way to store information of conditional relationships, compared to a complete CPT of all random variables. Look at the following example,

- B: Burglary occurs.
- A: Alarm goes off.
- E: Earthquake occurs
- J: John calls.
- M: Mary calls.



(a) How many CPTs do we need to store all the information, what are their size respectively?
(b) What is the size of the total CPT if we want to store probabilities of all possible results?
(c) write down the expression of $P(-b, -e, +a, +j, -m)$ using values in CPTs.
Hint: Use chain rule to break down the probability and use the fundamental assumption of BN to simplify the result.

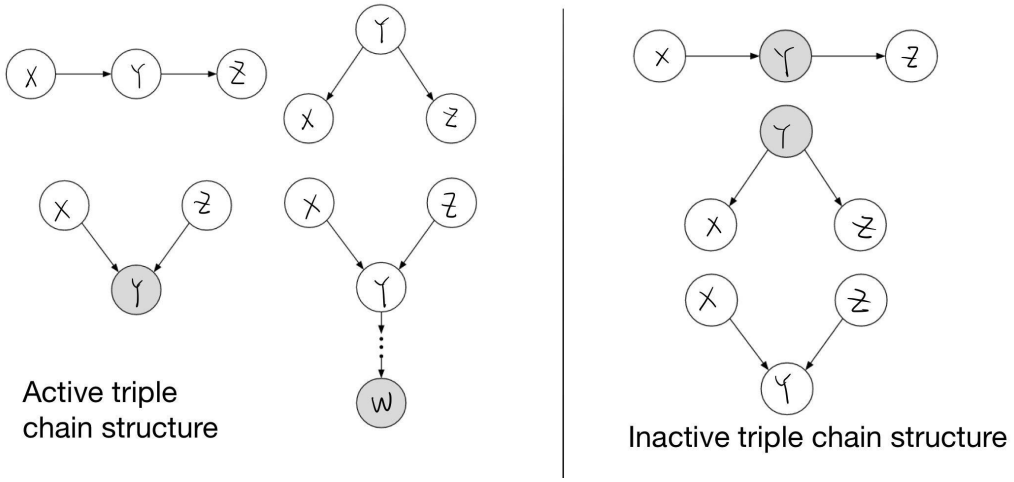
Solution 6. (a) 5 CPTs. with size (#cols * #rows, in order B, E, A, J, M) : $(0+2)*2^{0+1} = 4$, $(0+2)*2^{0+1} = 4$, $(2+2)*2^{2+1} = 32$, $(1+2)*2^{1+1} = 12$, $(1+2)*2^{1+1} = 12$.
(b) $(5+1)*2^5 = 192 > 4+4+32+12+12 = 64$.
(c) $P(-b, -e, +a, +j, -m) = P(-b) \cdot P(-e) \cdot P(+a \mid -b, -e) \cdot P(+j \mid +a) \cdot P(-m \mid +a)$
To conclude, given all of the CPTs for a graph, we can calculate the probability of a given assignment using the chain rule: $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents } (X_i))$.

Exercise 7 (Optional:Active & Inactive Triples of BN). Here we discuss the following BNs consisting of 3 nodes and 2 directed edges. They are called causal chain, common cause, common effect, and common effect with child observations. Also, we can decide whether Y is known to us (if it is known, in the figure, the corresponding node is shaded), so there are 4*2 = 8 cases. We are interested in either whether X and Z are independent ($X \perp\!\!\!\perp Z$), or whther X and Z are independent given the observation of Y ($X \perp\!\!\!\perp Z|Y$).

Show the following facts:

(a) In the triples in the left part of figure, X and Z are either not independent (Y not shaded), or not conditionally independent of Y (Y shaded), we call it **active triples**, because we can gain information about X via knowing Z and gain information about Z via knowing X.

(b) In the triples in the right part of figure, X and Z are either independent (Y not shaded), or conditionally independent of Y (Y shaded), we call it **inactive triples**, because we cannot gain any information about X via knowing X or gain any information about Z via knowing X.



Solution 7.