Introduction to Probability Theory – Part II [1]

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

[1] Largely based on material from Sharon Goldwater's tutorial *Basics of Probability Theory* (henceforth abbreviated as SGT), available at: `https://homepages.inf.ed.ac.uk/sgwater/math_tutorials.html`

Conditional Probability

The CONDITIONAL PROBABILITY of A given B, written $P(A|B)$, where the j is pronounced "given", is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

Example 4.1.1. Using the scenario from Exercise 2.8 of Section 2.5, with events A = "the student is male" and B = "the student is from the UK". What is $P(A|B)$?
Solution: $A \cap B$ is the set of male British students, so $P(A \cap B)$=$\frac{1}{7}$. $P(B) = \frac{4}{7}$, so $P(A \cap B) = \frac{1}{4}$.

Example 4.1.2. Using the same A and B as in Ex 4.1.1, what is $P(B|A)$?
Solution: We have the same $A \cap B$, so $P(A \cap B) = \frac{1}{7}$. $P(A) = \frac{3}{7}$, so $P(B|A) = \frac{\frac{1}{7}}{\frac{3}{7}} = \frac{1}{3}$

Product Rule

If we rearrange the terms in the definition of conditional probability (Eq 5),

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

we obtain the PRODUCT RULE:

$$P(A \cap B) = P(A|B)P(B) \tag{6}$$

This rule allows us to compute joint probabilities from conditional probabilities. Since intersection is commutative, we could also write:

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

Law of Total Probability Revisited

We can use the product rule in (6) to replace the joint probabilities in (4) with conditional probabilities and obtain the alternative equation in (7).

$$P(A \cap B) = P(A|B)P(B) \tag{6}$$

Let $\{E_1 \ldots E_n\}$ be a partition of S and $B \subseteq S$.

$$P(B) = \sum_{i=1}^{n} P(B \cap E_i) \tag{4}$$

$$P(B) = \sum_{i=1}^{n} P(B|E_i)P(E_i) \tag{7}$$

Chain Rule

The product rule can only handle the intersection of two events, but notice that we can apply it iteratively to expand the joint probability of several events into a product of several conditional probabilities. For example:

$$\begin{aligned} P(A \cap B \cap C \cap D) &= P(A|B \cap C \cap D)P(B \cap C \cap D) \\ &= P(A|B \cap C \cap D)P(B|C \cap D)P(C \cap D) \\ &= P(A|B \cap C \cap D)P(B|C \cap D)P(C|D)P(D) \end{aligned}$$

Chain Rule:

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = \prod_{i=1}^{n} P(E_i|E_{i+1} \cap \ldots \cap E_n) \qquad (8)$$

Chain Rule: Example 1

Drawing without replacement:

From a regular deck of 52 cards, evenly distributed among four suits, three cards are uniformly at random. What is the probability that these cards are all kings?

$$P(k1) \times P(k2|k1) \times P(k3|k1 \cap k2) = 4/52 \times 3/51 \times 2/50$$

Chain Rule: Example 2 [2]

Chain Rule:

$$P(E_1 \cap E_2 \cap ... \cap E_n) = \prod_{i=1}^{n} P(E_i | E_{i+1} \cap ... \cap E_n) \tag{8}$$

Word prediction task: n-gram models compute the probability of sequences as the product of subsequences:

$$\begin{aligned} p(\mathbf{w}) &= p(w_1, w_2, \ldots, w_M) \\ &= p(w_1) \times p(w_2 \mid w_1) \times p(w_3 \mid w_2, w_1) \ldots \\ &\quad \times p(w_M \mid w_{M-1}, \ldots, w_1) \end{aligned}$$

Each element in the product is the probability of a word given all of its preceding words.

[2] Example due to Eisenstein (2019). Introduction to Natural Language Processing, MIT Press, p. 123

Chain Rule: Example 2 (ctnd.)

The chain rule could also have been applied in reverse order.

$$p(\mathbf{w}) = p(w_M) \times p(w_{M-1} \mid w_M), \ldots, p(w_1 \mid w_2, \ldots, w_M)$$

Conditional probability distributions

$$\sum_{i=1}^{n} P(E_i|B) = \sum_{i=1}^{n} \frac{P(E_i \cap B)}{P(B)} \qquad \text{by defn of conditional prob.}$$

$$= \frac{1}{P(B)} \sum_{i=1}^{n} P(E_i \cap B)$$

$$= \frac{1}{P(B)} P(B) \qquad \text{by law of total probability}$$

$$= 1 \tag{9}$$

Conditional probability distributions

We also need to show that all the$P(E_i|B)$ are between 0 and 1. We know that P(B) and each $P(E_i \cap B)$ is between 0 and 1, because they are probabilities. Also,$P(E_i \cap B)$ cannot be larger than $P(B)$, because the intersection of B with another set can't be larger than B. But if $P(E_i \cap B) \leq P(B)$, and $P(B) > 0$ and $P(E_i \cap B) \geq 0$, then $\frac{P(E_i \cap B}{P(B)}$) must be between 0 and 1. Because a conditional probability distribution is a distribution, the rules for computing probabilities that we discussed already apply just as well to conditional distributions, provided we keep the conditioning event in all parts of the equation. In particular,

$$P(\neg E|B) = 1 - P(E|B) \tag{10}$$

$$P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C) \tag{11}$$

Conditional Probability Table (CPT)

- A CPT for the distribution conditioned on A has one column for each event E_i in the partition.
- Each column is labelled with E_i and lists the value of $P(E_i \mid A)$).

CPT: An example with letter bigrams[3]

	a	b	c	d	e	f	g	h
a	0.04	0.02	0.02	0.03	0.05	0.01	0.02	0.06
b	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01
c	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01
d	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.02
e	0.06	0.02	0.01	0.03	0.08	0.01	0.01	0.07
f	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01
g	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.02
h	0.08	0.00	0.00	0.01	0.10	0.00	0.01	0.02

[3] slide due to Cagri Cöltekin

CPT: Letter bigrams with marginal probabilities

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	
a	0.04	0.02	0.02	0.03	0.05	0.01	0.02	0.06	0.23
b	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04
c	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.05
d	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.08
e	0.06	0.02	0.01	0.03	0.08	0.01	0.01	0.07	0.29
f	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.02
g	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.07
h	0.08	0.00	0.00	0.01	0.10	0.00	0.01	0.02	0.22
	0.23	0.04	0.05	0.08	0.29	0.02	0.07	0.22	

JPT Example continued

Let us assume that probability of paying in-state and out-of state tuition is .55 and .45, respectively, and the probability of living on-campus and off-campus is .40 and .60, respectively. Then we can fill in the JPT as follows.

	on-campus	off-campus
P(in) = .55	P(in,on) = .22	P(in,off) = .33
P(out) = .45	P(out,on) = .18	P(out,off) = .27
	P(on) = .40	P(off) = .60

Then the probabilities to the left of each row and below each column are called marginal probabilities. Each marginal probability is the sum of all joint probabilities of the particular value for an attribute.

Independent Events

$$P(A|B) = P(A) \tag{12}$$

By substituting the left-hand side of Eq. (12) by Eq. (5) and multiplying each side by $P(B)$, we obtain Eq. (13):

$$P(A \cap B) = P(A)P(B) \tag{13}$$

Comment: If independence exists, the probability of each joint event in a JPT must be equal to the probabilities of the corresponding marginal probabilities. For example in JPT above: p(in,on) = p(in) $\times$ p(on) = .55 $\times$ .40 = .22

Example - due Khan Academy

Superpower	Male	Female	TOTAL
Fly	26	12	38
Invisibility	12	32	44
Other	10	8	18
TOTAL	48	52	100

1. Find the probability that the student chose to fly as their superpower: $P(Fly)$
2. Find the probability that the student was male: $P(Male)$
3. Find the probability that the student was male, given the student chose to fly as their superpower: $P(Male|Fly)$.

Example - due Khan Academy

Superpower	Male	Female	TOTAL
Fly	26	12	38
Invisibility	12	32	44
Other	10	8	18
TOTAL	48	52	100

1. Find the probability that the student chose to fly as their superpower, given the student was male: $P(Fly|Male)$
2. Is the following statement about conditional probability true or false? $P(A \mid B) = P(B \mid A)$. In general, you can reverse the order and the probability is the same either way
3. Interpret the meaning of $P(I \mid F) \approx 0.62$ and choose the correct paraphrase: 1. About 62 % of females chose invisibility as their superpower. 2. About 62 % of people who chose invisibility as their superpower were female.