Statistical Methods in Natural Language Processing (NLP)

Class 3: Conditional Probabilities

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Preceding Class: Recap

- we discussed some simple experiments.
- introduced notions, such as the *sample space*, denoted by S, as the set of possible outcomes of an experiment.

Preceding Class: Recap

- An experiment creates an outcome from a pre-determined set. For example, when we flip a coin once the outcomes is the set {head, tail} and when we roll a dice the outcomes in the set $\{1, 2, 3, 4, 5, 6\}$.
- An *event* as a set, namely, a collection of possible outcomes of an experiment ($E \subset S$).

Preceding Class: Recap

We introduced a number of fundamental concepts:

- unions ($A \cup B$),
- intersections ($A \cap B$),
- complements of events (A^c).

Preceding Class: Recap

- *Probability* is the measure of how likely an event is.
  - $P(A) = 0$: event A will almost definitely not occur.
  - $P(A)$ is close to 0.25: there is only a small chance that event A can occur.
  - $P(A) = 0.5$: there is a 50% chance that event A can occur.
  - P(A) is close to 0.75: there is a strong chance that event A can occur.
  - $P(A) = 1$: event A will almost definitely occur.

Preceding Class: Recap

- The probability of an event A happening is *the number of ways A can occur* divided by *the total number of possible outcomes*, if all outcomes are equally likely.
- The *marginal* or *unconditional probability* $P(A)$ is simply the probability of that A.
- The *joint probability* $P(A \cap B)$ or $P(A, B)$ has been defined as the probability of A and B.

Set operations in Python

1. `s.union(t)`
2. `s.intersection(t)`
3. `s.difference(t)`
4. `x in s`

Set operations in R

- `union(x, y)`
- `intersect(x, y)`
- `setdiff(x, y)`
- `setequal(x, y)`
- `is.element(el, set) or x %in% y.`

Today: An overview

- *Conditional Probability* $P(A|B) = P(A,B)/P(B)$, this is simply the probability of A, given that B occurred and is fundamental for understanding probability theory.
- The Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B.
- Any theorem that holds for probability also holds for conditional probability. We use this as the definition of conditional probability.
- We will introduce the *Bayes' Theorem*.

Kolmogorov axioms

In probability theory we can assume that for an event A in a sample space S, there is a value $P(A)$, which is the probability of A. The $P(A)$ follows the following three axioms:

First axiom: Non-negativity. The probability of an event is a non-negative real number:

$$0 \leq P(A) \leq 1$$

Second axiom: Normalization. The probability that at least one of the events in the entire sample space will occur is 1.

$$P(S) = 1$$

Third axiom: Finite additivity. Any countable sequence of disjoint sets (synonymous with mutually exclusive events) $A_1, A_2, \dots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

This suggests that, in a sequence of mutually exclusive events, the probability of at least one of these events to occur is equal to the sum of their probabilities.

Consequences

The probability axioms can lead to a number of important conclusions about probabilities. First, the second $1 = P(S)$ and third axiom $P(S) = P(A \cup A^c) = P(A) + P(A^c)$ can lead to the conclusion that the $1 = P(S) = P(A \cup Ac) = P(A) + P(A^c)$.

From this equation, we can conclude the following proposition:
$P(A^c) = 1 - P(A)$

The probability of the union of two events is equal to the probability of these events and the probability of their intersection. This principle is also known as the **inclusion-exclusion principle**.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

In other words, to calculate the union of A and B, we add the individual probabilities of A and BE and then subtract the intersection of A and B. So, if

A = {1, 2, 3, 4, 5} and B = {4, 5, 6, 7, 8}, the $A \cap B$ = {4, 5}, so the

$P(A \cup B)$ = {1, 2, 3, 4, 5} + {4, 5, 6, 7, 8} - {4, 5}
= {1,2,3,4,5,4,5,6,7,8} - {4, 5}
= {1,2,3,4,5,6,7,8}

we see that that if we had not done the subtraction of the second term, i.e., $P(A \cap B)$, the 4,5 would had appeared twice, which is not what we want.

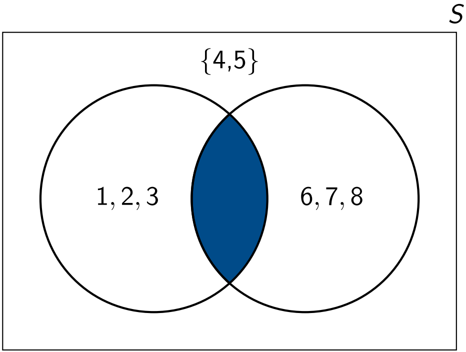


Figure: The shaded region represents intersection of the two events: $A \cap B$, which should be 4,5 not 4,5, 4,5.

Now for cases such $A = \{1, 2, 3, 4, 5\}$ and $B = \{6, 7, 8\}$, where the $A \cap B = \emptyset$, the result of the union is again $\{1, 2, 3, 4, 5, 6, 7, 8\}$ since we need only to add the two sets.

However, if there are three sets A, B, C, e.g.,

$$\begin{aligned}&\text{A} = \{1\ 2\ 3\ 4\ 5\ 6\}\\&\text{B} = \{5\ 6\ 7\ 8\ 9\ 10\}\\&\text{C} = \{1\ 2\ 5\ 6\ 7\ 11\}\end{aligned}$$

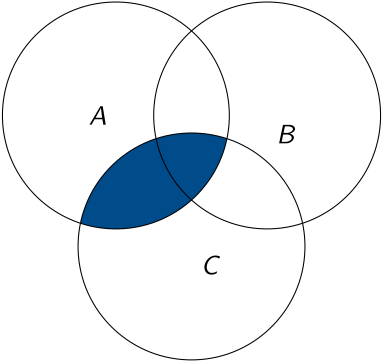


Figure: Shaded region: $A \cap C$.

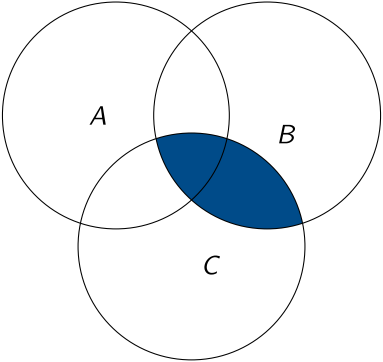


Figure: Shaded region: $B \cap C$.

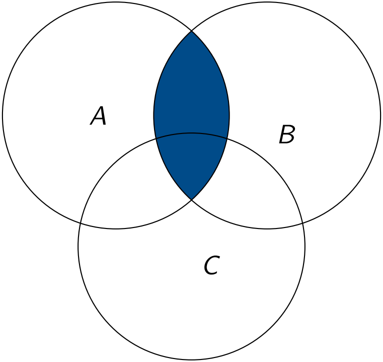


Figure: Shaded region: $A \cap C$.

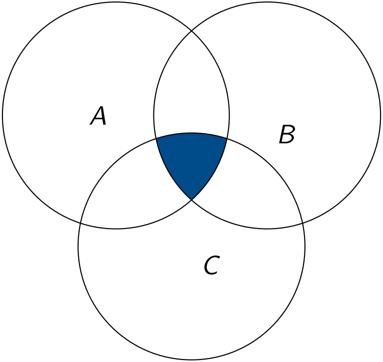


Figure: Shaded region: $A \cup B \cup C$.

$A \cup B \cup C$.

So again in this case, we need to add the individual probabilities $P(A)$, $P(B)$, $P(C)$ and subtract the intersections of these sets, namely, $(P(A \cap B) = 56$, $P(B \cap C) = 567$, and $P(A \cap C) = 1256$.

Nevertheless, in this case the calculation of $A \cup B \cup C$ should consider the intersection of the three sets ($A \cap B \cap C = 56$), which is deleted altogether because of the subtractions and add it back, so the final equation with the three sets A, B, and C is the following:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) -$$
$$P(A \cap B) - P(B \cap C) - P(A \cap C)$$
$$+ P(A \cap B \cap C)$$

In the more general case where there are n different sets A_i, the calculation of the unions of these sets is achieved using the following generic algorithm. So, if $A_1, \ldots, A_n$ are finite sets, then

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{i,j\,:\,1 \leq i < j \leq n} P(A_i \cap A_j) \\ + \sum_{i,j,k\,:\,1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(A_1 \cap \cdots \cap A_n).$$

In plain language, to calculate the number of elements in a finite union of finite sets, one should

1. take the sum of the individual sets.
2. subtract the number of elements, which appear in more than one set.
3. add back the number of elements which appear in more than two sets.
4. subtract the number of elements which appear in more than three sets.
5. . . .

The preceding discussion leads to the concept of **cardinality of a set**, which is the size of the set, namely how many elements are in a set. The cardinality of a set is denoted by $|A|$.

Counting equally probable events

What is the probability of an event E, that is a member of a set S of equally probable events? The answer is to this question is that the probability of any event E is equal to the proportion of outcomes in the S that are contained in E.

$$P(E) = \frac{number\ of\ events\ E}{number\ of\ S}$$

Let us consider this example: What is the probability of an event of a dice rolled twice to have a sum equal to 5?.

Answer

Remember that the sample space of two dice (i,j), where *i* is the first dice and *j* the second dice is 36. There are 4 possible outcomes, namely, (1,4), (2,3), (3,2), (4,1), so the probability is $P(5) = \frac{4}{36} = \frac{1}{9}$.

Two independent Events

Two random events are *independent* if knowing the value of one gives no information about the other. In other words two discrete A and B are independent if for *all* values of x and y

$$P(A = x, B = y) = P(A = x)P(B = y)$$

More formally, A and B (which have nonzero probability) are independent if the following is true:

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

More than two independent events

We can extend this idea to more than two events. Suppose that there are three events that are subset of S, i.e., $A, B, C, \subset S$, we say that A,B,C are mutually independent when

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

More than two independent events

For more than three sets $A_1, A_2 \ldots A_n \subset S$, then $A_1, A_2 \ldots A_n$ is mutually independent when

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1), P(A_2) \cap \ldots \cap (A_n)$$

The multiplication rule

Another useful idea is the multiplication rule, which states that when we multiply the probability of independent events we can find the probability of all events.

$$P(A_1 A_2 A_3 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \dots P(A_n|A_1 \dots A_{n1})$$

The multiplication rule: Proof

The multiplication rule is proven as follows:

$$P(A_1)\frac{P(A_1A_2)}{P(A_1)}\frac{P(A_1A_2A_3)}{P(E_1E_2)}\dots\frac{P(A_1A_2A\dots A_n)}{P(A1A2\dots A_{n-1})}=P(A_1A_2A_3\dots A_n)$$

Dependent Events

- An event can often be affected by the circumstances. For example, if you randomly select a word from a corpus of a language, the word class or part of speech can be any one of the part of speech in that language.
- However, if you know that the preceding word is an article then the probabilities to the get another article, or a verb are very small.

Conditional probability

- The conditional probability that the word we select from the corpus is a verb given that the previous word is an article is denoted by

$$P(verb|article)$$

Conditional probability

The conditional probability that an event A occurs given that event B has occurred is denoted by

$$P(A|B)$$

Conditional Probabilities

- **Conditional Probability:** is the probability of an event given that another event has occurred.

Dependent Events: Example

- If we roll two dice freely, the outcome can be one of the *36* possible outcomes, so the probability of an event happening in this case is $\frac{1}{36}$.
- If we want to get a sum of 5 and know that the first dice is 1, then there are only six possible outcomes for the second dice (1,1) (1,2) (1,3) (1,4) (1,5), and (1,6), so the probability to get the sum 5 is one $\frac{1}{6}$.

Dependent Events: Example

If the first dice is A and the second dice is B, we calculate their conditional probability as follows:

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

where $P(B) \neq 0$,

- The equation denotes that A is conditioned by B or in different words it denotes the conditional probability that an event A occurs **given** B.
- The conditional probability can be found by dividing the probability of the intersection of events $A \cap B$ by the probability of event A.
- So, the occurrence is a shared point in A and B, i.e., $P(A \cap B)$. Since A is given, B becomes the new sample space, which is of course smaller.

Example

If the first dice is A and the second dice is B, we calculate their conditional probability as follows: If A and B were *independent* then the equation would have been simply

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Second Example

If we flip a coin twice and all the occurrences in the sample space $S = (h, h), (h, t)(t, h), (t, t)$ are equally likely, what is the conditional probability that the coin in flip one and flip two lands on tails, given that flip one is a tail and?

- A = t,t: both tails
- B = (t,h), (t,t): the first coin be a tail.
- C = (t,h), (t,t), (h,t): at least one be a tail.

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(t, t)}{P((t, t), (t, h))} \\
&= \frac{1/4}{2/4} \\
&= \frac{1}{2}
\end{aligned}
$$

Second Example: Conditional probability

If we flip a coin twice and all the occurrences in the sample space $S = \{(h,h),(h,t)(t,h),(t,t)\}$ are equally likely, what is the conditional probability that the coin in flip one and flip two lands on tails, given that at least one flip is a tail?

Second Example: Conditional probability

$$\begin{aligned}
P(A|C) &= \frac{P(A \cap C)}{P(C)} \\
&= \frac{P(t,t)}{P((t,t),(t,h),(h,t))} \\
&= \frac{1/4}{3/4} \\
&= \frac{1}{3}
\end{aligned}$$

Law of Total Probability

Suppose that $A_1, A_2, A_3, \ldots A_n$ be a *partition* of the sample space S i.e., they are disjoint and their union is the entire sample space and B is an event of S, then

$$P(B) = \sum_{i_1}^{n} P(B|A_i)P(A_i)$$

Proof of the Law of Total Probability

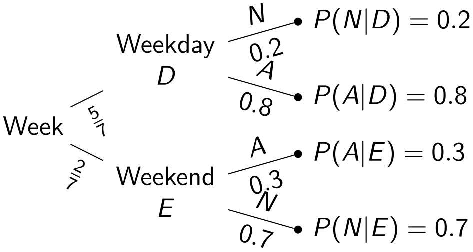
The proof can be defined as follows:

$$\begin{aligned} B &= (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_i) \\ &= P(B) = (B \cap A_1) + (B \cap A_2) + \ldots + (B \cap A_i) \\ &= P(B|A_1)P(A_1) + \ldots + P(B|A_n)P(A_n) \end{aligned}$$

Law of Total Probability: An example

In a city the probability for traffic accidents is 0.3 during the weekend and 0.2 during the other days of the week. **What is the overall probability for an accident happening from Monday to Sunday?**

- **Weekend**: E
- **Weekday**: D
- **No Accident**: N
- **Accident**: A



The probability of an accident during the whole week can be calculated from the total probability rule.

$$P(A) = P(A|E)P(E) + P(A|D)P(D) = 0.3 \times \frac{2}{7} + 0.2 \times \frac{5}{7} = 0.229$$

Bayes' theorem

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

where A and B are events and P(B) $\neq$ 0.

1. P(A) and P(B) are the independent probabilities of A and B respectively
2. $P(A|B)$, a conditional probability, is the probability of A given B.
3. $P(B|A)$ is the probability of B given A.

Bayes' Inference

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

where A and B are events and P(B) $\neq$ 0.

1. **H** a Hypothesis which we estimate from the evidence.
2. **E** new evidence that were not used in computing the prior probability.
3. **P(H)** is the **prior probability**: is the probability of a hypothesis before we evaluate it on a given evidence.
4. $P(H \mid E)$ is the **posterior probability**, this is the probability of H given E.
5. $P(E \mid H)$ is the **likelihood function** of the probability of E given a fixed H.
6. $P(E)$ a.k.a., the **marginal likelihood**.

Bayes' theorem

Note how the Bayes' Theorem differs from the conditional probability, which is

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

Bayes' Theorem

- The Bayes' Theorem has many applications in Computational Linguistics.
- It brakes the models in smaller parts.
- It is often used along with other machine learning applications, like the HMMs.

Normalizing Constant

- The denominator $P(B)$ is known as a normalizing constant.
- We can omit it, if we want to find out which event from a set is more possible, given A:

$$\arg\min_B = \frac{P(B \mid A)\, P(A)}{P(B)} = \arg\min_B = P(B \mid A)\, P(A) \tag{1}$$

Bayes' theorem

So let us assume again that in a city the probability for traffic accidents is 0.3 during the weekend and 0.2 during the other days of the week. This time we want to know **what is the probability of being a weekend day E given an accident A**.

Bayes' theorem: Example

$$\begin{aligned}
P(E \mid A) &= \frac{P(A \mid E)P(E)}{P(A)} \\
&= \frac{0.3 \times \frac{2}{7}}{P(A) = P(A|E)P(E) + P(A|D)P(D) = 0.3 \times \frac{2}{7} + 0.2 \times \frac{5}{7}} \\
&= \frac{0.086}{0.229} = 0.375
\end{aligned}$$

So the probability of being a Weekend given an Accident is 0.375.

Next Class

Distributions and Random Variables

1. Discrete Variables
2. Continuous Variables
3. Distributions
4. Bernoulli Distribution
5. Binomial Distribution
6. Hypergeometric Distribution