

Independence

Based on a chapter by Chris Piech and Lisa Yan

Pre-recorded lecture: Sections 1 and 2.

In-lecture: Section 3: DeMorgan’s Laws.

Covered later: Section 4 (as part of pre-lecture Lecture 6)

1 Independence

Independence is a big deal in machine learning and probabilistic modeling. Knowing the “joint” probability of many events (the probability of the “and” of the events) requires exponential amounts of data. By making independence and conditional independence claims, computers can essentially decompose how to calculate the joint probability, making it faster to compute, and requiring less data to learn probabilities.

Independence

Two events, E and F , are **independent** if and only if:

$$P(EF) = P(E)P(F)$$

Otherwise, they are called **dependent** events.

This property applies regardless of whether or not E and F are from an equally likely sample space and whether or not the events are mutually exclusive.

The independence principle extends to more than two events. In general, n events E_1, E_2, \dots, E_n are independent if for every subset with r elements (where $r \leq n$) it holds that:

$$P(E_{i_1}, E_{i_2}, \dots, E_{i_r}) = P(E_{i_1})P(E_{i_2}) \dots P(E_{i_r})$$

The general definition implies that for three events E, F, G to be independent, *all* of the following must be true:

$$P(EFG) = P(E)P(F)P(G) \tag{1}$$

$$P(EF) = P(E)P(F) \tag{2}$$

$$P(EG) = P(E)P(G) \tag{3}$$

$$P(FG) = P(F)P(G) \tag{4}$$

Example: Dice

Roll two fair, 6-sided dice. Let event E be where the first die outcome is 1, event F be where the second die outcome is 6. Note that E and F are independent because $P(E) = 1/6$, $P(F) = 1/6$, and $P(EF) = 1/36$, where $EF = \{(1, 6)\}$.

Problem 1 : Define G as the event where the sum of the two dice rolls is 5. Are E and G independent?

Solution 1 : No; E and G are dependent. Note that $P(G) = \frac{4}{36} = \frac{1}{9}$, because there are 4 ways to roll a sum of 5: $\{(1, 4), (2, 3), (3, 2), (4, 1)\}$. However, the event EG only has one outcome: $\{(1, 4)\}$, and thus $P(EG) = \frac{1}{36}$. Since $P(EG) \neq P(E)P(G)$, E and G are dependent.

Problem 2 : Define G as the event where the sum of the two dice rolls is 7. Are E , F , and G independent?

Solution 2 : No; E , F , and G are dependent. To prove independence of three events E , F , and G , we must show the four equalities above (Equations (1)-(4)). If any of these fail, then the events are dependent.

Note that $P(G) = \frac{6}{36} = \frac{1}{6}$, because there are 6 outcomes in $G = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$. The event $EFG = \{(1, 6)\}$, and thus $P(EFG) = \frac{1}{36}$. Since $P(EFG) \neq P(E)P(F)P(G)$, E , F , and G are dependent.

This example demonstrates how pairwise independence does not imply independence of three events. It can be shown that $P(EF) = P(E)P(F)$, $P(EG) = P(E)P(G)$, and $P(FG) = P(F)P(G)$, yet this is not sufficient to prove independence of E , F , and G together.

2 Independent trials

Problems with more than two independent events come up frequently. For example: the outcomes of n separate flips of a coin are all independent of one another. Each flip in this case is called an **independent trial** of the experiment, where the probabilities of events associated with each flip are identical.

Example: Genetics

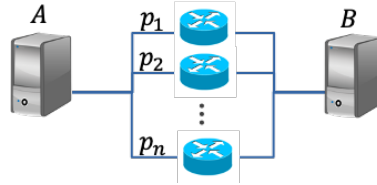
Problem: Two parents both have an (A, a) gene pair for some trait. Each parent will pass on their genes (each gene equally likely) to their child. Define a child having a recessive trait as the child receiving the gene pair (a, a) (i.e., one a gene from each parent), and suppose this happens with probability 0.25. Suppose the parents have three children, where each child is an independent trial. What is the probability that all three children have the recessive trait?

Solution: Let E_i be the event that child i has the recessive trait, where $P(E_i) = 0.25$ and $i = 1, 2, 3$. We would like to calculate $P(E_1E_2E_3)$, the probability that all three children have the recessive trait. Using the chain rule,

$$\begin{aligned} P(E_1E_2E_3) &= P(E_1)P(E_2|E_1)P(E_3|E_1E_2) \\ &= P(E_1)P(E_2)P(E_3) = (0.25)^3, \end{aligned}$$

where because E_1, E_2, E_3 are independent, $P(E_2|E_1) = P(E_2)$ and $P(E_3|E_1E_2) = P(E_3)$. In general, the probability of the intersection (“AND”) of independent events is simply the product of the probabilities of the independent events, individually. In this example, we explicitly showed this simplification using the chain rule, but the statement of independence is sufficient in most cases.

Example: Network Reliability



Problem: Consider the parallel network shown above, where there are n links between A and B , and each link has its own router. These routers function independently, where each router has probability p_i of functioning, where $1 \leq i \leq n$. Let E be the event where there exists a functional path from A to B . What is $P(E)$?

Solution: The event E where a functional path exists is equivalent to the event where at least one router works. We note that E is the complement of E^C , the event where no routers work. Let F_i be the event of router i functioning for $i = 1, \dots, n$, where $P(F_i) = p_i$.

$$\begin{aligned}
 P(E) &= P(F_1 \cup F_2 \cup \dots, \cup F_n) = 1 - P(F_1^C \cap F_2^C \cap \dots \cap F_n^C) \\
 &= 1 - (1 - p_1)(1 - p_2)(1 - p_n) = 1 - \prod_{i=1}^n (1 - p_i).
 \end{aligned}$$

The first line is an application of DeMorgan’s Law, which we discuss further in Section 3.

Example: Flipping a Biased Coin

A biased coin is flipped n times. Each flip (independently) comes up heads with probability p , and tails with probability $1 - p$. What is the probability of getting exactly k heads?

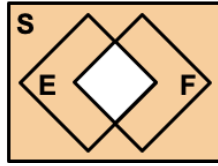
Solution: Consider all the possible orderings of heads and tails that result in k heads. There are $\binom{n}{k}$ such orderings, and all of them are mutually exclusive. Since all of the flips are independent, to compute the probability of any one of these orderings, we can multiply the probabilities of each of the heads and each of the tails. There are k heads and $n - k$ tails, so the probability of each ordering is $p^k(1 - p)^{n-k}$. Adding up all the different orderings gives us the probability of getting exactly k heads: $\binom{n}{k}p^k(1 - p)^{n-k}$.

(Spoiler alert: This is the probability density of a **binomial distribution**. Intrigued by that term? Stay tuned for next week!)

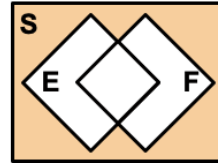
3 DeMorgan’s Laws

In the same way that the mutual exclusion property makes it easier to calculate the probability of the OR of two events, independence makes it easier to calculate the AND of two events.

DeMorgan’s Laws are a pair of laws from set theory that transform unions into intersections, and vice-versa, using set complements. For two events E and F , these laws can be illustrated as in Figure 1 and written formally as follows:



$$(E \cap F)^C = E^C \cup F^C$$



$$(E \cup F)^C = E^C \cap F^C$$

Figure 1: DeMorgan's Laws for two events E and F in sample space S .

DeMorgan's Laws

For events E_1, E_2, \dots, E_n ,

$$\left(\bigcap_{i=1}^n E_i \right)^C = \bigcup_{i=1}^n E_i^C$$

$$\left(\bigcup_{i=1}^n E_i \right)^C = \bigcap_{i=1}^n E_i^C$$

In probability, this implies the following:

$$P(E_1 E_2 \cdots E_n) = 1 - P(E_1^C \cup E_2^C \cup \cdots \cup E_n^C)$$

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = 1 - P(E_1^C E_2^C \cdots E_n^C)$$

These laws can be applied to probability of events to apply the properties of mutual exclusion and independence, where appropriate. The next set of examples illustrate this flexibility (where in Hash Tables Part C, the probabilities are quite complicated since the resulting events are neither mutually exclusive nor independent).

Example: Hash Map

Let's consider our friend the hash map. Suppose m strings are hashed (unequally) into a hash table with n buckets. Each string hashed is an independent trial, with probability p_i of getting hashed to bucket i . Calculate the probability of these three events:

- A) $E =$ the first bucket has ≥ 1 string hashed to it
- B) $E =$ at least 1 of buckets 1 to k has ≥ 1 string hashed to it
- C) $E =$ each of buckets 1 to k has ≥ 1 string hashed to it

Part A

Let S_i be the event that string i is hashed into the first bucket. Note that all S_i are independent of one another. The complement, S_i^C , is the event that string i is not hashed into the first bucket; by mutual exclusion, $P(S_i^C) = 1 - p_1 = p_2 + p_3 + \dots + p_n$.

$P(E) = P(S_1 \cup S_2 \cup \dots \cup S_m)$	Definition of S_i
$= 1 - P((S_1 \cup S_2 \cup \dots \cup S_m)^C)$	Complement
$= 1 - P(S_1^C S_2^C \dots S_m^C)$	De Morgan's Law
$= 1 - P(S_1^C)P(S_2^C) \dots P(S_m^C)$	since the events are independent
$= 1 - (1 - p_1)^m$	calculating $P(S_i)$ by mutual exclusion

Part B

Let F_i be the event that at least one string is hashed into bucket i . Note that the F_i 's are neither independent nor mutually exclusive.

$P(E) = P(F_1 \cup F_2 \cup \dots \cup F_k)$	
$= 1 - P([F_1 \cup F_2 \cup \dots \cup F_k]^C)$	since $P(A) + (A^C) = 1$
$= 1 - P(F_1^C F_2^C \dots F_k^C)$	by De Morgan's law
$= 1 - (1 - p_1 - p_2 - \dots - p_k)^m$	mutual exclusion, independence of strings

The last step is calculated by realizing that $P(F_1^C F_2^C \dots F_k^C)$ is only satisfied by m independent hashes into buckets other than 1 through k .

Part C

Let F_i be the same as in Part B.

$P(E) = P(F_1 F_2 \dots F_k)$	
$= 1 - P([F_1 F_2 \dots F_k]^C)$	since $P(A) + P(A^C) = 1$
$= 1 - P(F_1^C \cup F_2^C \cup \dots \cup F_k^C)$	by De Morgan's (other) law
$= 1 - P\left(\bigcup_{i=1}^k F_i^C\right)$	
$= 1 - \sum_{r=1}^k (-1)^{r+1} \sum_{i_1 < \dots < i_r} P(F_{i_1}^C F_{i_2}^C \dots F_{i_r}^C)$	by General Inclusion/Exclusion

where $P(F_1^C F_2^C \dots F_k^C) = (1 - p_1 - p_2 - \dots - p_k)^m$ just like in the last problem.

Example: Simplified Craps

Two 6-sided dice are rolled repeatedly. Consider the sum of the two dice. What is $P(E)$, where E is defined as the event where a sum of 5 is rolled before a sum of 7?

Define our independent trials to be each paired roll. We can then define F_i as the event where we observe our first 5 before 7 on the n -th trial. In other words, no 5 or 7 was rolled in the first $n - 1$ trials, and a sum of 5 was rolled on the n -th trial. Notice that the F_i for $i = 1, \dots, \infty$ are mutually exclusive, as there is only ever one first occurrence of the sum of 5. The probability of rolling a sum of 5 or a sum of 7 is $\frac{4}{36}$ and $\frac{6}{36}$, respectively, and therefore $P(F_i) = (\frac{26}{36})^{n-1}(\frac{4}{36})$.

$$\begin{aligned}
 P(E) &= P(F_1 \cup F_2 \cup \dots \cup F_i \cup \dots) = \sum_{i=1}^{\infty} P(F_i) && F_i \text{ mutually exclusive} \\
 &= \frac{4}{36} \sum_{i=1}^{\infty} (\frac{26}{36})^{i-1} = \frac{4}{36} \sum_{i=0}^{\infty} (\frac{26}{36})^i \\
 &= \frac{4}{36} \cdot \frac{1}{1 - \frac{26}{36}} = \frac{2}{5}
 \end{aligned}$$

where the last line comes from the property of infinite geometric series, where $|x| < 1$: $\sum_{i=1}^{\infty} x_i = \frac{1}{1-x}$.

4 Conditional Independence

Two events E and F are called **conditionally independent** given a third event G , if

$$P(EF | G) = P(E | G)P(F | G)$$

Or, equivalently:

$$P(E | FG) = P(E | G)$$

Conditioning Breaks Independence

An important caveat about conditional independence is that ordinary independence does not imply conditional independence, nor the other way around.

Knowing when exactly conditioning breaks or creates independence is a big part of building complex probabilistic models; the first few weeks of CS 228 are dedicated to some general principles for reasoning about conditional independence. We will talk about this in another lecture. I included an example in this handout for completeness:

Example: Fevers

Let's say a person has a fever if they either have malaria or have an infection. We are going to assume that getting malaria and having an infection are independent: knowing if a person has malaria does not tell us if they have an infection. Now, a patient walks into a hospital with a fever. Your belief that the patient has malaria is high and your belief that the patient has an infection is high. Both explain why the patient has a fever.

Now, given our knowledge that the patient has a fever, gaining the knowledge that the patient has malaria *will* change your belief the patient has an infection. The malaria explains why the patient has a fever, and so the alternate explanation becomes less likely. The two events (which were previously independent) are dependent when conditioned on the patient having a fever.

Example: Faculty Night

At faculty night with a CS professor in attendance, you observe 44 students. Of these, you find out that 30 are straight-A students. Additionally, 20 of the 44 are CS majors, and of these 20, 6 are straight-A students.

Let A be the event that a student gets straight A's, C be the event that a student is a CS major, and F be the event that a student attends faculty night. In probability notation, $P(A|F) = 30/44 \approx 0.68$, but $P(A | C, F) = 6/20 = 0.30$. It would seem that being a CS major decreases your chance of being a straight-A student!

You decide to investigate further by surveying your whole dorm. There are 100 students in your dorm; 30 of these are straight-A students, 20 are CS majors, and 6 are straight-A CS majors. That is, overall, $P(A) = 30/100 = 0.30$, and $P(A | C) = 6/20 = 0.30$. So A and C are independent! What happened at faculty night?

As it turns out, faculty night attracted two types of people: straight-A students (who go to all the faculty nights), and CS majors. So the non-straight-A students at this faculty night are more likely to be CS majors! It's not because CS students are slackers, or because CS is harder; it's because non-straight-A students with other majors didn't come to faculty night.

In both of these examples, conditioning on an event E leads to dependence between previously independent events A and B when A and B are independent causes of E .