Chris Piech
CS 109

Lecture Notes #5
October 3, 2018

Independence

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 $E_1, E_2, \ldots, E_n$ are independent if for every subset with r elements (where $r \leq n$) it holds that:

$$P(E_a, E_b, \ldots, E_r) = P(E_a)P(E_b)\ldots P(E_r)$$

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

$$\begin{aligned} P(EFG) &= P(E)P(F)P(G) \\ P(EF) &= P(E)P(F) \\ P(EG) &= P(E)P(G) \\ P(FG) &= P(F)P(G) \end{aligned}$$

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 a "trial" of the experiment.

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.

Example 1: 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!)

2 Analytic Probability

At this point in your probability journey, you know many of the core rules of analytic probability. That means you have a lot of tools at your disposal. Even if you understand how each tool works, you still need to learn how to approach a problem, and how to select the right tool. Here is a picture which briefly summarizes the many approaches we have talked about so far:

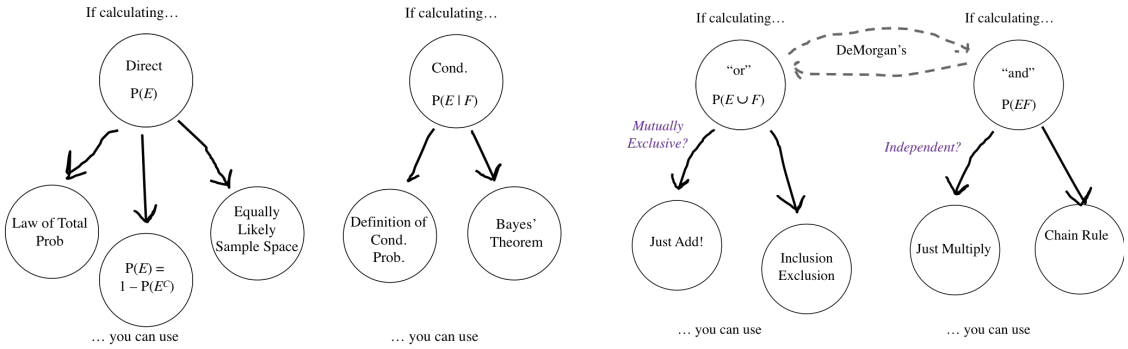

Figure 1: Analytic Probability Tools

We have introduced two properties for pairs of events: mutual exclusion and independence. Remember that if events are mutually exclusive, it is easy to compute the “or" of the events. If events are independent, it is easy to compute the “and" of events. Sometimes you will be explicitly told you can assume that a particular group of events are either independent or mutually exclusive. Other times you can make a logical argument that they are (for example if you are hashing strings into a hashmap, the event that a string hashes to one bucket is mutually exclusive from the event that it hashes to another).

You can use complements and De-Morgan’s law to turn computing probability with “and" into probability with “or" (and vice versa). This can be helpful in situations where, for example, you are asked to calculate the probability of the “or" of two events that are independent.

De Morgan's Law for Probability For any two events E and F:

$$P((E \cap F)^C) = P(E^C \cup F^C) \qquad \text{Version 1}$$
$$P((E \cup F)^C) = P(E^C \cap F^C) \qquad \text{Version 2}$$

This often is used in conjunction with the rule that the sum of an event and its complement is 1:

$$\begin{aligned} P(E \cap F) &= 1 - P((E \cap F)^C) && \text{Since } P(E) + P(E^C) = 1 \\ &= 1 - P(E^C \cup F^C) && \text{By DeMorgan's law} \\ P(E \cup F) &= 1 - P((E \cup F)^C) && \text{Since } P(E) + P(E^C) = 1 \\ &= 1 - P(E^C \cap B^F) && \text{By DeMorgan's law} \end{aligned}$$

Finally, there are often many ways to solve a problem. Try and come up with more than one. In all cases sanity check your answers. All probabilities should be positive and less than or equal to 1.

Example 2: 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 F_i be the event that string i is not hashed into the first bucket. Note that all F_i are independent of one another. By mutual exclusion, $P(F_i) = (p_2 + p_3 + \cdots + p_n)$.

$$\begin{aligned} P(E) &= 1 - P(E^C) && \text{since } P(A) + (A^C) = 1 \\ &= 1 - P(F_1 F_2 \ldots F_m) && \text{definition of } F_i \\ &= 1 - P(F_1)P(F_2) \ldots P(F_m) && \text{since the events are independent} \\ &= 1 - (p_2 + p_3 + \cdots + p_n)^m && \text{calculating } P(F_i) \text{ by mutual exclusion} \end{aligned}$$

Part B

Let Fi 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.

$$\begin{aligned}
P(E) &= P(F_1 \cup F_2 \cup \cdots \cup F_k) \\
&= 1 - P([F_1 \cup F_2 \cup \cdots \cup F_k]^C) && \text{since } P(A) + (A^C) = 1 \\
&= 1 - P({F_1}^C {F_2}^C \ldots {F_k}^C) && \text{by De Morgan's law} \\
&= 1 - (1 - p_1 - p_2 - \cdots - p_k)^m && \text{mutual exclusion, independence of strings}
\end{aligned}$$

The last step is calculated by realizing that $P({F_1}^C {F_2}^C \ldots {F_k}^C)$ 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.

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

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

3 Conditional Independence

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

$$P(EF \mid G) = P(E \mid G)P(F \mid G)$$

Or, equivalently:

$$P(E \mid FG) = P(E \mid G)$$

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 3: 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.