David Varodayan
CS109

Lecture Notes #5
January 15, 2020

Independence

Based on a chapter by Chris Piech

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!)

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 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, $P(S_i^C) = 1 - p_1 = p_2 + p_3 + \cdots + p_n$.

$$
\begin{aligned}
P(E) &= P(S_1 \cup S_2 \cup \cdots \cup S_m) && \text{Definition of } S_i \\
&= 1 - P((S_1 \cup S_2 \cup \cdots \cup S_m)^C) && \text{Complement} \\
&= 1 - P(S_1^C S_2^C \ldots S_m^C) && \text{De Morgan's Law} \\
&= 1 - P(S_1^C)P(S_2^C) \ldots P(S_m^C) && \text{since the events are independent} \\
&= 1 - (1 - p_1)^m && \text{calculating } P(S_i) \text{ by mutual exclusion}
\end{aligned}
$$

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.

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:

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

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 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.

Example 4: 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 $P(A \mid C, F) = 6/20 = 0.30$. 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 $P(A \mid C) = 6/20 = 0.30$. 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.