CS 30: Discrete Math in CS (Winter 2019): Lecture 23

Date: 18th February, 2019 (Monday)

Topic: Probability: Conditional Independence and Bayes Rule

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please discuss in Piazza/email errors to deeparnab@dartmouth.edu

---

1. **Some Recap.**

   - $\mathbf{Pr}[\mathcal{A} \mid \mathcal{B}] = \frac{\mathbf{Pr}[\mathcal{A} \cap \mathcal{B}]}{\mathbf{Pr}[\mathcal{B}]}$.
   - $\mathcal{A}$ and $\mathcal{B}$ are *independent* events if $\mathbf{Pr}[\mathcal{A} \cap \mathcal{B}] = \mathbf{Pr}[\mathcal{A}] \cdot \mathbf{Pr}[\mathcal{B}]$.
   - If $\mathbf{Pr}[\mathcal{B}] \neq 0$, then an equivalent way of stating independence is $\mathbf{Pr}[\mathcal{A} \mid \mathcal{B}] = \mathbf{Pr}[\mathcal{A}]$.

2. **Mutual and Pairwise Independence.** Say $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are three events. These are said to be *pairwise* independent, if any pair of them are independent events. That is,

   $$\mathbf{Pr}[\mathcal{A} \cap \mathcal{B}] = \mathbf{Pr}[\mathcal{A}] \cdot \mathbf{Pr}[\mathcal{B}], \quad \mathbf{Pr}[\mathcal{B} \cap \mathcal{C}] = \mathbf{Pr}[\mathcal{B}] \cdot \mathbf{Pr}[\mathcal{C}], \quad \mathbf{Pr}[\mathcal{C} \cap \mathcal{A}] = \mathbf{Pr}[\mathcal{C}] \cdot \mathbf{Pr}[\mathcal{A}]$$

   These three events are *mutually independent* if these are pairwise independent *and*

   $$Pr[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}] = \mathbf{Pr}[\mathcal{A}] \cdot \mathbf{Pr}[\mathcal{B}] \cdot \mathbf{Pr}[\mathcal{C}]$$

   ✍

   > **Exercise:** *Describe three events which are* pairwise independent *but not* mutually *independent.*

   More generally, given events $A_1, A_2, \ldots, A_n$ are mutually independent, if for any two *disjoint* subsets S and T of $\{1, 2, \ldots, n\}$, the event $\cap_{i \in S} A_i$, that is the intersection of all events indexed by S, and the event $\cap_{j \in T} A_j$ are independent. ✍

   > **Exercise:** *Describe three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ such that $\mathbf{Pr}[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}] = \mathbf{Pr}[\mathcal{A}] \cdot \mathbf{Pr}[\mathcal{B}] \cdot \mathbf{Pr}[\mathcal{C}]$ but they are not pairwise independent (and therefore not mutually independent).*

3. **Conditional Independence.**

   Consider the following two events. There lies in front of you a *fair* coin. Alice tosses it. Then Bob tosses the same coin. Let $\mathcal{A}$ be the event that Alice gets heads. Let $\mathcal{B}$ be the event that Bob gets heads. Are these independent? Even before doing the calculation, you would say sure. Alice's toss shouldn't hinder Bob's toss. Indeed, both $\mathbf{Pr}[\mathcal{A}] = \mathbf{Pr}[\mathcal{B}] = 1/2$ and $\mathbf{Pr}[\mathcal{A} \cap \mathcal{B}] = 1/4$. These are independent. ✍

   > **Exercise:** *Check that $\mathcal{A}$ and $\mathcal{B}$ are independent even when the coin is not* fair*, but instead it came heads all the time, or came heads 90% of the time.*

Now consider a slightly different experiment. In a box lies two coins. One is fair. The other is biased; it comes heads with probability 0.75. You pick up a coin from these two at random and place it in front of you. Alice tosses it. Bob tosses the same coin. $\mathcal{A}$ and $\mathcal{B}$ are same as above. Are these independent events?

To see (or at least get a hunch – in Math you should always have a hunch) that they are not before doing any calculations, take the experiment to an extreme. Suppose both the coins in the box were super un-fair; suppose one of them came tails all the time, and the other came heads all the time. Then note, if $\mathcal{A}$ occurs, then $\mathcal{B}$ occurs with 100% probability (if Alice sees a head, then she has for sure picked the all-heads coin, and so Bob will for sure see a heads as he is tossing the same coin). On the other hand, $\mathcal{B}$ is not a sure-shot; if I had picked the all-tails coin, then $\mathcal{B}$ doesn't occur. Thus, $\mathcal{A}$ and $\mathcal{B}$ aren't independent.

However, there is a *third* random event here. It is the event $\mathcal{E}$ which is whether I pick the fair coin or not. I claim that $\mathcal{A}$ and $\mathcal{B}$ are independent *if we condition on* $\mathcal{E}$. That is, I claim

$$\mathbf{Pr}[\mathcal{A} \cap \mathcal{B} \mid \mathcal{E}] = \mathbf{Pr}[\mathcal{A} \mid \mathcal{E}] \cdot \mathbf{Pr}[\mathcal{B} \mid \mathcal{E}]$$

Indeed, if I tell you that $\mathcal{E}$ has occurred, then the problem becomes the one asked before; given a fair coin tossed by Alice and Bob, the events that they see heads is independent.

Remark:

Conditional Independence is a tricky concept. Be wary. Here are a couple of plausible potholes.

- $\mathcal{A}$ and $\mathcal{B}$ are independent events. Then they are also *conditionally* independent on any event $\mathcal{E}$. **False.** *Example: Roll two fair dice. $\mathcal{A}$ is the event that the first dice is odd. $\mathcal{B}$ is the event that the second dice is odd. These are independent events. Now consider the event $\mathcal{E}$ that the* sum of the two dice is odd.. *What is $\mathbf{Pr}[\mathcal{A} \mid \mathcal{E}]$? You can now calculate this – it is $1/2$ as well. Similarly, $\mathbf{Pr}[\mathcal{B} \mid \mathcal{E}] = 1/2$. However, what is $\mathbf{Pr}[\mathcal{A} \cap \mathcal{B} \mid \mathcal{E}]$? Yep, it's zero.* ***Independence can be lost upon conditioning.***
- $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{E}$. Then they are conditionally independent given $\neg\mathcal{E}$ as well. **False.** *In its generality this is false, although in the above example of coins, it is true. To see why it is false, we can consider again the setting of rolling two dice. However, this time $\mathcal{A}$ occurs if the first die lands 1, and $\mathcal{B}$ occurs if the second die lands 1. $\mathcal{E}$ is the event that the sum is 2; $\neg\mathcal{E}$ is the event that the sum is not 2.*

  Note: $\mathbf{Pr}[\mathcal{A} \mid \mathcal{E}] = \mathbf{Pr}[\mathcal{B} \mid \mathcal{E}] = \mathbf{Pr}[\mathcal{A} \cap \mathcal{B} \mid \mathcal{E}] = 1$. Thus, $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{E}$. On the other hand, $\mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{E}]$ is something non-zero (figure out what it is!), and $\mathbf{Pr}[\mathcal{B} \mid \neg\mathcal{E}]$ is something non-zero. But, $\mathbf{Pr}[\mathcal{A} \cap \mathcal{B} \mid \neg\mathcal{E}]$ is certainly zero. ***Conditional Independence can be lost upon the negation of the event we are complementing on.***

4. **Bayes Rule.** Put simply, Bayes' rule is the following observation: for any two events $\mathcal{A}$ and $\mathcal{B}$ which each occur with non-zero probability

$$\mathbf{Pr}[\mathcal{B} \mid \mathcal{A}] = \frac{\mathbf{Pr}[\mathcal{A} \mid \mathcal{B}] \cdot \mathbf{Pr}[\mathcal{B}]}{\mathbf{Pr}[\mathcal{A}]} \qquad \text{(Bayes Rule)}$$

The proof is trivial after we substitute the formula of conditional probability.

We can expand it slightly more using the law of total probability to get

$$\mathbf{Pr}[\mathcal{B} \mid \mathcal{A}] = \frac{\mathbf{Pr}[\mathcal{A} \mid \mathcal{B}] \cdot \mathbf{Pr}[\mathcal{B}]}{\mathbf{Pr}[\mathcal{A} \mid \mathcal{B}] \cdot \mathbf{Pr}[\mathcal{B}] + \mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{B}] \cdot (1 - \mathbf{Pr}[\mathcal{B}])} \qquad \text{(Bayes Rule - Opened up)}$$

Why is this a big deal? We will look at three examples. But in a nutshell, it states that to answer what is the probability of event $\mathcal{B}$ given event $\mathcal{A}$, if we know (a) the total probability of event $\mathcal{B}$, and if (b) the probability of event $\mathcal{A}$ is *easier* to figure out, then we can get our answer. The main applications come in when $\mathcal{A}$ is an "outcome" and $\mathcal{B}$ is a "hypothesis"; $\mathbf{Pr}[\mathcal{B}]$ is a "*prior* belief" on the hypothesis, and $\mathbf{Pr}[\mathcal{B} \mid \mathcal{A}]$ is our "*posterior* belief" given we see the outcome event $\mathcal{A}$.

Examples.

- *Arithmophobia is a quality of life debilitating condition and should be detected as early as possible. Fortunately, the pharmaceutical company* HAYSTEAM *have come up with a test. It's not perfect. It has a* ***false positive rate*** *of* $\mathsf{fp} = 1\%$*; that is, on* 1% *of the healthy population, the test detects the condition, It also has a* ***false negative rate*** *of* $\mathsf{fn} = 2\%$*. Again, this means that* 2% *of the afflicted population go undetected. It is assumed around* 40% *of the population may be suffering from Arithmophobia.*

  You take the test and unfortunately it comes positive (the test says you have Arithmophobia). What is the probability that you actually do?

  What a story! But such situations abound. Easy-peasy if you know Bayes rule and know how to set things up.

  $\mathcal{A}$ be the event that you have the affliction. Now, you don't know whether you do or not (that's why, presumably, you take the test). Before taking the test, you just look at the statistics and believe that you are as likely as anyone else to have this condition. Since 40% of the population have it, you conclude (reasonably)

  $$\mathbf{Pr}[\mathcal{A}] =: p_A = 0.4$$

  $\mathcal{P}$ be the event that the test comes out positive on you. We are really interested in figuring out $\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}]$. We will do so by applying Bayes rule.

  First, is $\mathbf{Pr}[\mathcal{P} \mid \mathcal{A}]$ easy? That is, if you did have the affliction, what is the probability that the test would catch it? The answer is $(1 - \mathsf{fn})$; you would be tested positive unless we got a *false negative*. Thus,

  $$\mathbf{Pr}[\mathcal{P} \mid \mathcal{A}] = (1 - \mathsf{fn}) = 0.98$$

  How about $\mathbf{Pr}[\mathcal{P} \mid \neg\mathcal{A}]$? This is precisely the *false positive* rate. So,

  $$\mathbf{Pr}[\mathcal{P} \mid \neg\mathcal{A}] = \mathsf{fp} = 0.01$$

  Now to apply Bayes rule,

  $$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}] = \frac{\mathbf{Pr}[\mathcal{P} \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P} \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}] + \mathbf{Pr}[\mathcal{P} \mid \neg\mathcal{A}] \cdot (1 - \mathbf{Pr}[\mathcal{A}])}$$

which simplifies to

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}] = \frac{(1-\mathsf{fn})p_A}{(1-\mathsf{fn})p_A + \mathsf{fp}(1-p_A)} \;=\; 0.985$$

Thus, if the test comes positive, then the chances you have the affliction gets close to 98.5%.

It is instructive to repeat the above calculations when $\mathbf{Pr}[\mathcal{A}]$ is small, say $\mathbf{Pr}[\mathcal{A}] = 0.1$, that is, only 10% of the population have Arithmophobia. In that case, we get $\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}] = \frac{(0.98)(0.1)}{(0.98)(0.1)+(0.01)(0.9)} = 0.915$. After the positive test, my belief that I have the affliction goes from 10% to more than 90%.

If $\mathbf{Pr}[\mathcal{A}] = 1\%$, however, then if you repeat the calculation you get that $\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}] = 0.497$. Thus, if the affliction is so rare that less than 1% of the people have the disease, then a positive test (with the given rates) shouldn't take your belief to more than a random (fair) coin toss.

More generally, to have "high belief probabilities" the false-negative and false-positive scores should be *substantially* smaller than the (prior) probability of the affliction (hypothesis) itself.

- *Spam Filters. We are trying to train a (Bayesian) Spam Filter. We start with a corpus with* 2000 *spam messages and* 1000 *real messages. We observe that the word "Congratulations" appears in* 100 *spam messages, and* 10 *real messages. We also observe that the word "Account" appears in* 160 *spam messages and* 20 *real messages. Assume you believe that any incoming email is possible spam with probability* 40%*. What is the probability an incoming message is spam given it contains the word "Congratulations"? What is the probability an incoming message is spam given it contains the word "account"? What is the probability that the incoming message is spam, given it contains* **both** *words "account" and "congratulations"? If we set a threshold of* 90% *to mark spam or not, in which of these cases would we mark spam.*

  Remark: *I couldn't cover this example in class. It is an interesting and informative example. In fact, before moving ahead, please try it yourself.*

Consider an incoming email. Let $\mathcal{S}$ be the event that it is spam. The assumption we are making is that $\mathbf{Pr}[\mathcal{S}] = 0.4$.

Let $\mathcal{A}$ be the event that the word "account" appears in the email. Let $\mathcal{C}$ be the event that the word "congratulations" appears in the email. From the data, we *deduce* that in a random spam message, the chances of seeing "congratulations" is $\frac{100}{2000} = 0.05$. Thus, we conclude

$$\mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] = 0.05$$

Similarly, we conclude,

$$\mathbf{Pr}[\mathcal{C} \mid \neg\mathcal{S}] = \frac{10}{1000} = 0.01$$

since $\neg\mathcal{S}$ implies a 'real' message. Also, we conclude

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] = \frac{160}{2000} = 0.08$$

and

$$\mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{S}] = \frac{20}{1000} = 0.02$$

Now, we can apply Bayes rule to get

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{A}] = \frac{\mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}]}{\mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}] + \mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{S}] \cdot \mathbf{Pr}[\neg\mathcal{S}]} = \frac{(0.08) \cdot (0.4)}{(0.08)(0.4) + (0.02)(0.6)}$$

which computes to 0.727. That is, if we see the word "account" in an incoming mail, we would believe the probability it is spam is around 72.7%. Thus, out spam-filter won't mark it spam.

Similarly, for "congratulations", we get

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{C}] = \frac{\mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}]}{\mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}] + \mathbf{Pr}[\mathcal{C} \mid \neg\mathcal{S}] \cdot \mathbf{Pr}[\neg\mathcal{S}]} = \frac{(0.05) \cdot (0.4)}{(0.05)(0.4) + (0.01)(0.6)}$$

which computes to around 0.769. That is, if we see the word "account" in an incoming mail, we would believe the probability it is spam is around 77%. The spam-filter won't mark this spam.

How do we solve the next question – when we see both "congratulations" and "account". Well, we need to find

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{A} \cap \mathcal{C}] = \frac{\mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}]}{\mathbf{Pr}[\mathcal{A} \cap \mathcal{C}]} \tag{1}$$

We **don't know** how to calculate $\mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \mathcal{S}]$. This is where (another) assumption, called the **Naive Bayes Assumption** is made. In the setting of Spam Filters, it states that the events $\mathcal{A}$ and $\mathcal{S}$ are *conditionally independent* on *both* spam (that is $\mathcal{S}$) and *real* messages. What it says that it does recognize that the distribution of these words ("congratulations", "account") may not behave independently on the whole email corpus; but if we focus our attention to the classes at hand, then it does. Again, this is an *assumption*, which is actually made out there many time in the real world.

$$\mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \mathcal{S}] = \mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \mathcal{S}], \quad \mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \neg\mathcal{S}] = \mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \neg\mathcal{S}] \quad \text{(Naive Bayes)}$$

Once we make it, then our calculations can start again. We get:

$$\mathbf{Pr}[\mathcal{A} \cap \mathcal{C}] = \mathbf{Pr}[\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \mathcal{S}] + \mathbf{Pr}[\neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \neg\mathcal{S}]$$

and the RHS, with the Naive Bayes assumption, becomes

$$\mathbf{Pr}[\mathcal{A} \cap \mathcal{C}] = \mathbf{Pr}[\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] + \mathbf{Pr}[\neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \neg\mathcal{S}]$$

Substituting in the Bayes rule formula (1), we get

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{A} \cap \mathcal{C}] = \frac{\mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}]}{\mathbf{Pr}[\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \mathcal{S}] + \mathbf{Pr}[\neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{A} \mid \neg\mathcal{S}] \cdot \mathbf{Pr}[\mathcal{C} \mid \neg\mathcal{S}]}$$

which evaluates to

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{A} \cap \mathcal{C}] = \frac{(0.05)(0.08)(0.4)}{(0.05)(0.08)(0.4) + (0.02)(0.01)(0.6)} = 0.9302$$