CS 30: Discrete Math in CS (Winter 2020): Lecture 15-Supplement

Date: 5th February, 2020 (Wednesday)
Topic: Probability: Conditional Independence
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. **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 and tosses 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 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, none of the events individually is a sure-shot. 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. The events $\mathcal{A}$ and $\mathcal{B}$ are therefore *independent conditioned on the event $\mathcal{E}$.*

Remark: *Conditional Independence is a tricky concept. Be wary. For example:*

- "$\mathcal{A}$ 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 die is odd. $\mathcal{B}$ is the event that the second die 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 conditioning on.***

2. **Revisiting the "Two-Tests" example.** Suppose $\mathcal{A}$ is your initial belief you have an affliction (based on, say, statistics). There is a test which has a false negative rate of fn and a false positive rate of fp. That is, if you have the affliction, the probability the test says you don't is fn, and if you do not have the affliction, the probability the test says you do is fp. You take the test once and see a positive. You take the test again and you see a positive. What are the chances you do have the affliction. This is a problem we solved using Bayes law. And we saw there were "two ways" to do this.

   Say $\mathcal{P}_1$ is the event the first test comes positive. $\mathcal{P}_2$ is the event the second test comes positive. After one test, the probability we do have the affliction is

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1] = \frac{\mathbf{Pr}[\mathcal{P}_1 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_1]} \tag{1}$$

   and, the probability we do have an affliction *after* two tests coming positive is

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1, \mathcal{P}_2] = \frac{\mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2]} \tag{2}$$

   Now note, crucially, that $\mathcal{P}_1$ and $\mathcal{P}_2$ are **not independent**. Much like the example above of the coin being pulled out of a bag in the previous bullet point. However, they are *conditionally independent* on both $\mathcal{A}$ and $\neg\mathcal{A}$. That is,

$$\mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2 \mid \mathcal{A}] = \mathbf{Pr}[\mathcal{P}_1 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \quad \text{and} \quad \mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2 \mid \neg\mathcal{A}] = \mathbf{Pr}[\mathcal{P}_1 \mid \neg\mathcal{A}] \cdot \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}]$$

   In particular, this implies

$$\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}, \mathcal{P}_1] = \mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \qquad \text{and} \qquad \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}, \mathcal{P}_1] = \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}]$$

   Where are we getting at? Well, now we can "simplify" (2) as

$$\begin{aligned}
\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1, \mathcal{P}_2] &= \frac{\mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_1, \mathcal{P}_2]} \\
&= \frac{(\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{P}_1 \mid \mathcal{A}]) \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{P}_1] \cdot \mathbf{Pr}[\mathcal{P}_1]} \qquad \text{Cond. Indep.} \\
&= \frac{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{P}_1]} \cdot \left(\frac{\mathbf{Pr}[\mathcal{P}_1 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A}]}{\mathbf{Pr}[\mathcal{P}_1]} \right)
\end{aligned}$$

Note that the paranthesized expression is precisely, by (1), $\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1]$. Thus, we get

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1, \mathcal{P}_2] = \frac{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1]}{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{P}_1]} \tag{3}$$

Now, by the law of total probability,

$$\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{P}_1] = \mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}, \mathcal{P}_1] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1] + \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}, \mathcal{P}_1] \cdot \mathbf{Pr}[\neg\mathcal{A} \mid \mathcal{P}_1]$$

and by conditional independence, we get

$$\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{P}_1] = \mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1] + \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}] \cdot (1 - \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1])$$

Substituting in (3), we get

$$\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1, \mathcal{P}_2] = \frac{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1]}{\mathbf{Pr}[\mathcal{P}_2 \mid \mathcal{A}] \cdot \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1] + \mathbf{Pr}[\mathcal{P}_2 \mid \neg\mathcal{A}] \cdot (1 - \mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1])}$$

which is exactly what you would have if only $\mathcal{P}_2$ occurred with the prior $\mathbf{Pr}[\mathcal{A}]$ changed to $\mathbf{Pr}[\mathcal{A} \mid \mathcal{P}_1]$.

3. **An example with Bayes rule and Conditional Independence**

   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.*

   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

$$\mathbf{Pr}[\mathcal{S} \mid \mathcal{A} \cap \mathcal{C}] = \frac{\mathbf{Pr}[\mathcal{A} \cap \mathcal{C} \mid \mathcal{S}] \cdot \mathbf{Pr}[\mathcal{S}]}{\mathbf{Pr}[\mathcal{A} \cap \mathcal{C}]} \tag{4}$$

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.

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

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 (4), 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

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