Lisa Yan
CS 109

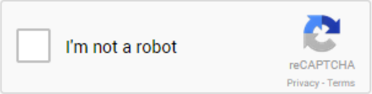
Section #5
May 13-15, 2020

Section 5

Based on the work of many CS109 staffs

1. **Quiz 2 is Coming.** Quiz 2 is less than a week away, and once again, section will be optional. But as your dedicated CS109 staff, we refuse to let the good times in section stop! So we wanted to know, when would work best for you?

2. **ReCaptcha.** Based on browser history, Google believes that there is a 0.2 probability that a particular visitor to a website is a robot. They decide to give the visitor a reCaptcha:



Google presents the visitor with a 10 mm by 10 mm box. The visitor must click inside the box to show that they are not a robot. You have observed that robots click uniformly in the box. However, the distance location of a human click has X location (mm from the left) and the Y location (mm from the top) distributed as independent normals both with mean $\mu = 5$ and variance $\sigma^2 = 4$.

   a. What is the probability density function of a robot clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

   b. What is the probability density function of a human clicking $X = x$ mm from the left of the box and $Y = y$ mm from the top of the box?

   c. The visitor clicks in the box at ($x = 6$ mm, $y = 6$ mm). What is Google's new belief that the visitor is a robot?

a.

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{100} & \text{if } 0 \le x, y \le 10 \\ 0 & \text{else} \end{cases}$$

b.

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x) f_Y(y) && \text{independence} \\ &= \frac{1}{(2\sqrt{2\pi})^2} e^{-\frac{(x-5)^2}{8}} e^{-\frac{(y-5)^2}{8}} && \text{normal PDF} \\ &= \frac{1}{8\pi} e^{-\frac{(x-5)^2}{8}} e^{-\frac{(y-5)^2}{8}} && \text{normal PDF} \end{aligned}$$

c. Let Click be the event that the user clicked at location ($x = 6$ mm, $y = 6$ mm). We can then use Bayes rule (with law of total probability in the denominator):

$$
\begin{aligned}
P(\text{Robot}|\text{Click}) &= \frac{f(\text{Click}|\text{Robot})P(\text{Robot})}{f(\text{Click})} \\
&= \frac{f(\text{Click}|\text{Robot})P(\text{Robot})}{f(\text{Click}|\text{Robot})P(\text{Robot}) + f(\text{Click}|\text{Human})P(\text{Human})} \\
&= \frac{\frac{1}{100} \cdot 0.2}{\frac{1}{100} \cdot 0.2 + \frac{1}{8\pi} e^{-\frac{(1)^2}{8}} e^{-\frac{(1)^2}{8}} \cdot 0.8} \approx 0.075
\end{aligned}
$$

3. **Binary Tree**: Consider the following function for constructing binary trees:

```
struct Node {
    Node *left;
    Node *right;
};

Node *randomTree(float p) {
    if (randomBool(p)) {  // returns true with probability p
        Node *newNode = new Node;
        newNode->left = randomTree(p);
        newNode->right = randomTree(p);
        return newNode;
    } else {
        return nullptr;
    }
}
```

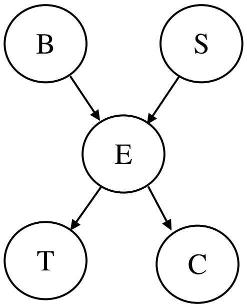
The `if` branch is taken with probability p (and the `else` branch with probability $1 - p$). A tree with no nodes is represented by `nullptr`; so a tree node with no left child has `nullptr` for the `left` field (and the same for the right child).

Let X be the number of nodes in a tree returned by `randomTree`. You can assume $0 < p < 0.5$. What is $E[X]$, in terms of p?

Let X_1 and X_2 be number of nodes the left and right calls to `randomTree`.
$E[X_1] = E[X_2] = E[X]$.

$$\begin{aligned}
E[X] &= p \cdot E[X \mid \texttt{if}] + (1-p)E[X \mid \texttt{else}] \\
&= p \cdot E[1 + X_1 + X_2] + (1-p) \cdot 0 \\
&= p \cdot (1 + E[X] + E[X]) \\
&= p + 2pE[X] \\
(1-2p)E[X] &= p \\
E[X] &= \frac{p}{1-2p}
\end{aligned}$$

4. **Monitoring Satellites**. As a part of CS109's Aeronautics and Space Administration agency, we are in charge of monitoring satellites. Recently, we have had several issues with a satellite; due to our probability expertise, we are in charge of modeling what has happened. As a first step, we decide to model our satellite using a Bayesian Network and binary variables.



Using our background knowledge, we decide to focus on trajectory deviation (T) and communication loss (C). We know that both of these issues are a result of electric system failure (E), which can result from either battery failure (B) or solar panel failure (S). For each variable, let 1 correspond to the failure occurring, and 0 otherwise.

a. What is the probability that everything is working?

For the sake of more manageable notation, we will write $X = 0$ as X_0.

$$P(T_0, C_0, E_0, S_0, B_0) = P(B_0)P(S_0)P(E_0|B_0, S_0)P(T_0|E_0)P(C_0|E_0)$$

We can think of this as the “chain rule" for Bayesian Networks.

b. What is the probability that there is a battery failure, if we have observed trajectory deviation and communication loss?

By definition, we know that

$$P(B_1|T_1,C_1) = \frac{P(B_1,T_1,C_1)}{P(T_1,C_1)}.$$

For the numerator, by chain rule, we know that

$$P(B_1,T_1,C_1) = \sum_{i=0}^{1}\sum_{j=0}^{1} P(B_1)P(S_i)P(E_j|B_1,S_i)P(T_1|E_j)P(C_1|E_j).$$

For the denominator, also by chain rule, we know that

$$P(T_1,C_1) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} P(B_k)P(S_i)P(E_j|B_k,S_i)P(T_1|E_j)P(C_1|E_j).$$

This is similar to what we saw in lecture, except with extra sums.

c. What is the probability that the electric system fails?

We can use the same idea from (b) to solve this problem.

$$P(E_1) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{l=0}^{1} P(B_i)P(S_j)P(E_1|B_i,S_j)P(T_k|E_1)P(C_l|E_1).$$

From here, we can use our bare hands to brute force a solution. However, calculating a quadruple sum feels bad, and if there were more nodes, or our random variables had a larger range, well, we would be in trouble. Thus, by being clever, we can use the structure in a Bayesian Network to make our work easier. Intuitively, T and C does not affect E, so it feels like we should be able to ignore them. But does the math work out that way? Starting from our equation above, we can re-arrange our sums to obtain

$$P(E_1) = \sum_{i=0}^{1}\sum_{j=0}^{1} P(B_i)(S_j)P(E_1|B_i,S_j)(\sum_{k=0}^{1}\sum_{l=0}^{1} P(T_k|E_1)P(C_l|E_1)).$$

This may not look like we have achieved anything, but the key observation is that

$$\sum_{k=0}^{1}\sum_{l=0}^{1} P(T_k|E_1)P(C_l|E_1) = 1$$

so we can safely drop the third and fourth summation. Without drowning ourselves in details, we can think about just communication loss and electrical system failure, where it is more clear that

$$\sum_{l=0}^{1} P(C_l|E_1) = 1.$$

We can then conclude that

$$P(E_1) = \sum_{i=0}^{1} \sum_{j=0}^{1} P(B_i)P(S_j)P(E_1|B_i, S_j).$$

In other words, our intuition is correct — we can just ignore the T and C, and thus work with less variables. This is one way to think about a problem “graphically", instead of “mathematically". Sometimes, thinking about numbers is hard, and thinking about graphs is easier.

d. Let us assume now that we believe that there is an electrical system failure and a battery failure. How does that change whether or not we think that there is a solar panel failure? (Conceptual, for fun!)

We may be tempted to think that, because B and S do not have any shared ancestry, they are independent. But this is not true. Intuitively, if we believe that there is a battery failure, then it reduces our belief that there is a solar panel failure because we have already found an explanation for our electric system failure. Stated even more informally, if we have already explained something (“the electric system failure was caused by a battery failure"), we do not go searching for more explanations.

If we wanted a more mathematical explanation, we can use our bag of tricks to find $P(S_1|E_1)$, and $P(S_1|E_1, B_1)$; we should find that $P(S_1|E_1) \geq P(S_1|E_1, B_1)$.

This is just to show another way that Bayesian Networks capture common reasoning patterns. If they are intuitive to us, that is because they were meant to be! However, it also shows that we have to be careful when we move away from numbers and into graphs. For example, if we do not have any information about our electric system, i.e. we calculate $P(S_1)$, not $P(S_1|E_1, B_1)$, then S and B are independent — which is what we would expect. In some sense, the issue with the scenario is that S and B are connected via E; but if they are not, then they are in fact independent.

As an aside, we do not need to think about T and C, for the same reasons as part (c).