Machine Learning 10-701 due October 3, 2013

http://alex.smola.org/teaching/cmu2013-10-701x/

Carnegie Mellon University

Homework 1 Solutions

---

1 Probability Review [Ahmed; 20pts]

1.1 Why just 2 variables ? Let's go for 3 [6 pts]

(a)

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} = \frac{P(A,B|C)P(C)}{P(B|C)P(C)} = \frac{P(A,B|C)}{P(B|C)}$$

(b)

$$P(A|C) = \frac{P(A,C)}{P(C)} = \sum_B \frac{P(A,B,C)}{P(C)} = \sum_B P(A,B|C)$$

(c) By applying (b) then (a)

$$P(A|C) = \sum_B P(A,B|C) = \sum_B P(A,B|C)P = \sum_B P(A|B,C)P(B|C)$$

1.2 Evaluating Test Results [8 pts]

Let $A1$, $A2$ and S denote the events of using algorithm 1, using algorithm 2 and success of the used algorithm respectively.

- $$P(A2|S) = \#(\text{A2 used and succeeded})/\#(\text{used algorithm succeeded}) = 2150/(2150+6000) \approx 0.264$$

- We care about the success rate of each algorithm

  $$P(S|A1) = \#(\text{A1 used and succeeded})/\#(\text{A1 used}) = 6000/(6000+1700) \approx 0.779$$
  $$P(S|A2) = \#(\text{A2 used and succeeded})/\#(\text{A2 used}) = 2150/(2150+500) \approx 0.811$$

  Since $P(S|A2) > P(S|A1)$ then using $A2$ is recommended (if success rate is the only aspect we care about).

  Common Mistakes: Some students were confused by the fact that the algorithms were tested with sets of data that significantly differ in size. This should not be a concern as long as (1) the two sets are drawn from the same distribution [And that's true because the decision to use one of the algorithms was independent of the transaction] and (2) the data is sufficient to confirm the *statistical significance* of the result (i.e. state that, with a high confidence, the difference in success rates is not a product of chance). If you are concerned about data size, then you need to establish a statistical test or confidence intervals (some of you actually did that) though it was not required but the mere fact that A2 was tested using less examples should not an issue by itself. In fact, it is typical to test A2 with far less examples (may be 1%) so that we don't risk our business to test a new method.

- We define X_1 to be the event that "$A1$ will succeed *if* applied on the transaction" and similarly define X_2 for $A2$.

Then we know from the previous part that

$$P(X_1) = P(S|A_1) = 0.779$$
$$P(X_2) = P(S|A_2) = 0.811$$

The claim states that

$$P(X_2|X_1) = 0.7$$

Using the law of total probability

$$P(X_2) = P(X_2|X_1)P(X_1) + P(X_2|\neg X_1)P(\neg X_1)$$
$$P(X_2|\neg X_1) = \frac{P(X_2) - P(X_2|X_1)P(X_1)}{P(\neg X_1)} = \frac{0.811 - 0.7 * 0.779}{1 - 0.779} = 1.202 > 1$$

Then the claim cannot be true.

There are many derivations that share the same idea: showing that this claim violates probability axiom. For example, you can work out the joint distribution of X_1 and X_2 and show that it does not sum to 1. Another more insightful way is assume that $P(X_2|\neg X_1) = 1$ (i.e. $A2$ succeeds in all cases where $A1$ fails), which gives you the minimum possible value of $P(X_2|X_1)$ that can achieve the known success rate $P(X_2)$. This minimum value turns out to be approximately 0.76. In other words, the claimed value is too small that even if $A2$ succeeded in all $A1$ failure cases, it would be still not sufficient to achieve its known success rate.

Common Mistakes:

- Some students answered the question this way: since this probability cannot be computed using the data we have then the claim cannot be true. This shows that *the data is not sufficient to compute the claimed quantity* but you were asked to show that *given the data, this claim is impossible*.
- Many students did not realize that we are talking about a new set of events X_1, X_2 which are not related to the events used in the previous parts except by the fact that $P(X_i) = P(S|A_i)$ and hence they attempted to solve the question in terms of A_i and S, which lead to wrong arguments. For example, $P(A_2|A_1)$ is simply zero because we never use both algorithms on the same transaction in the scenario from which we derived the A_i events.
- Some of you claimed that X_1 and X_2 are independent and argued that $P(X_2|X_1) = P(X_2)$ and that's not true; in the extreme case, X_1 and X_2 can be identical algorithms or algorithms that make opposite predictions.

1.3 Monty Hall Problem [6pts]

$$\begin{aligned}
\frac{P(car3|choose1, open2)}{P(car1|choose1, open2)} &= \frac{P(car3, open2|choose1)/P(open2|choose1)}{P(car3, open2|choose1)/P(open2|choose1)} \\
&= \frac{P(car3, open2|choose1)}{P(car3, open2|choose1)} \\
&= \frac{P(open2|car3, choose1)P(car3|choose1)}{P(open2|car1, choose1)P(car1|choose1)}
\end{aligned}$$

Homework 1 Solutions

By the game rules, the host cannot open the chosen door or the door containing the car. If two doors are valid, the door to be opened is chosen uniformly at random. This gives:

$$P(open2|car3, choose1) = 1$$
$$P(open2|car3, choose1) = 1/2$$

Also the door to place the car behind is chosen at random and independent of the player's choice. So

$$P(car1|choose1) = P(car1) = P(car3) = P(car3|choose1)$$

Substitution gives

$$\frac{P(car3|choose1, open2)}{P(car1|choose1, open2)} = 1/0.5 = 2$$

Common Mistakes:

- The most common mistake is attempting to compute $P(choose1)$ or in general $P(choose1, ...|...)$. The two typical answers were $P(choose1) = 1/3$ and $P(choose1) = 1$. $P(choose1)$ is unknown because we don't know the player's strategy, that's why we keep conditioning on it. $P(choose1) = 1$ does not mean *Assume the player chooses door1*, which is correctly expressed by conditioning on $choose1$. It means that the player is known to always choose door 1 in each game.
- Some students enumerated all possible choose1,openX,carY with their probabilities and concluded that "the car is twice more likely to be in the closed door that was not chosen". However, by marginalizing over all possible $openX$ rather than conditioning on $open2$, you assume that the player does *not* know which door is opened. This is a different problem even though it happens to have the same solution (because of the way the host chooses the door to open). Without a clear argument that the two problems have similar properties, the solution is incomplete.
- Another "mistake" is attempting to calculate the value of common factors that will cancel out anyway when taking the ratio (e.g. $P(open2|choose1)$). It is not a real mistake because the solution is still correct (and so it didn't affect your grade), but it is a waste of time and I am not talking about the extra few minutes you needed for this assignment. I am talking about potentially several hours or more of computation time if you unnecessarily compute normalization constants when implementing an algorithm.