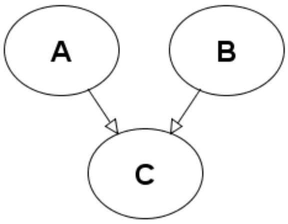
**A short story illustrating why pure machine learning (without expert input) may be doomed to fail and totally unnecessary**

***Norman Fenton*[1]*, Queen Mary University of London and Agena Ltd,  19 Nov 2012***

A and B are two medical conditions very well known to doctors Bill and Ludmila. They both know that these conditions are pretty rare (both have an incidence of about one in 1,000 people). There is a third medical condition C (whose name is "FiroziliRalitNoNeOba") that Bill has heard the name of, but knows nothing about. However, he *has* heard that patients with either A or B usually also have C. Bill has a massive database of 600,000 people with the details of which conditions they have. The data looks like this:

| Patient number | A | B | C |
|---|---|---|---|
| 1 | No | No | No |
| 2 | No | No | No |
| 3 | Yes | No | Yes |
| 4 | No | No | No |
| 5 | No | No | No |
| 6 | No | No | No |
| 7 | Yes | No | Yes |
| 8 | No | Yes | Yes |
| 9 | No | No | No |
| 10 | No | No | No |
| 11 | No | No | No |
| 12 | No | Yes | Yes |
| 13 | No | No | No |
| 14 | No | No | No |
| | …. | … | … |
| | …. | … | … |
| 600,000 | No | No | No |

He feels he can use this database to 'discover' the underlying causal model (Bayesian Network) relating A, B, and C. So he approaches Fred - a machine learning specialist who says he can easily learn the true relationship from the data. Ludmila says this is a waste of time. She says that her expertise about all three diseases enables her to produce the correct Bayesian Network (BN) without data. She says the BN structure looks like this:



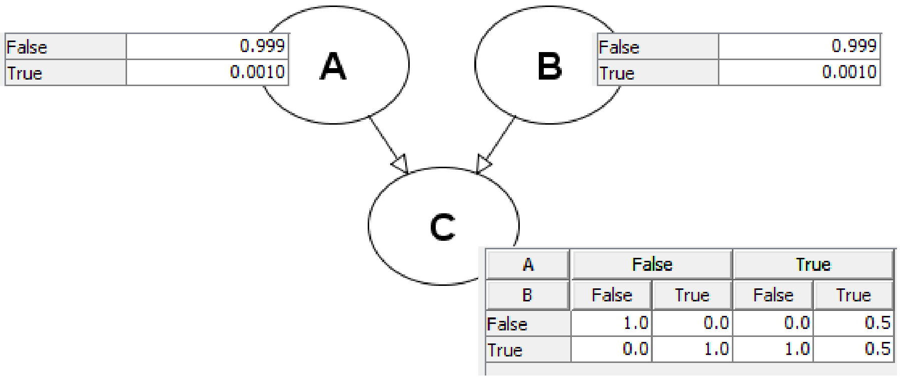

She says she can even have a very good stab at writing down the full probability table for C (i.e. the probability of C given the different combinations of states of A and B).

Citing well known studies about doctors' poor expert judgement involving subjective probabilities, Fred warns Bill that relying on Ludmila's expert judgement is both dangerous and unnecessary; unlike the expert, the learning algorithm will produce the true unbiased model. He explains why the algorithm starts with no prior assumptions about relationships between the nodes. It will discover them from the data.

[1] norman@eecs.qmul.ac.uk

Of the 600,000 people in the dataset there are 600 people with condition A, 600 with condition B and 1200 with condition C. Every single person with condition A also has C and every single person with B also has C.

Fred proudly presents the learnt Bayesian Network model to Bill and Ludmila:




| A | |
|---|---|
| False | 0.999 |
| True | 0.0010 |

| B | |
|---|---|
| False | 0.999 |
| True | 0.0010 |

| A | False | | True | |
|---|---|---|---|---|
| B | False | True | False | True |
| False | 1.0 | 0.0 | 0.0 | 0.5 |
| True | 0.0 | 1.0 | 1.0 | 0.5 |

Fred says: “Ludmila, I am pleased to say that your expert judgement about the structure was exactly right. And of course we have also learnt the correct probabilities.”

“But”, Ludmila says, “the probability table for C is wrong. I do not agree with the last column. Why are those probabilities equal?”

“Ummm….”, says Fred, “…that’s because the dataset was not big enough.”

“But you had the full data for 600,000 people and you only needed to learn 3 Boolean variables”, says Ludmila.

Fred responds: “Ah, but in none of those 600,000 entries did we find even one instance of a person who had both A and B, so we could not learn anything about C when both A and B are present. After all, because A and B each have an incidence of 1 in a 1000 – and because we found absolutely no direct relationship between them - there is actually a better than evens chance that you would not find this combination in 600,000 people. So we really did not have any relevant data and the 0.5 probabilities are what our algorithm produces in this case.”

At this point Bill chips in: “Well surely if patients with A ***always*** have C and if patients with B ***always*** have C then why can’t we simply conclude that C must be true when both A and B are.”

“Well” says Fred, “that sounds reasonable, but that would mean using some expert subjective judgement.

“Well”, says Ludmila, “I can PROVE that the learnt model is wrong AND that the ‘reasonable’ solution offered by Bill is also wrong. The name of Condition C - FiroziliRalitNoNeOba - is actually a Russian word. Its literal translation is ‘*A person suffering from either Firoz or Ralit but not both*’. ‘Firoz’ is the Russian word for condition A and ‘Ralit’ is the Russian word for condition B. I told you I knew everything about this condition and could provide you with the correct probability table but you did not trust me.”

***Moral of the story:***

- Sometimes you have to trust experts to provide a far more informed quantitative judgement than you will get from data alone.
- Even really big datasets will be insufficient for some very small problems.
- Trusting the expert can save you a whole load of unnecessary data-collection and machine learning effort.