Bayesian Argumentation

Stephan Hartmann

Munich Center for Mathematical Philosophy

Workshop “Modèles de la Rationalité”
Paris, December 2014

Motivation

- Goal: Study argumentation from a Bayesian point of view.
- Examine different argument types: deduction, induction, inference to the best explanation (IBE), the no-alternatives argument (NAA),...
- Here I will focus on an analysis of deductive inference patterns such as *modus ponens*.

Overview

1. The Main Idea
2. The Kullback-Leibler Divergence and Conditionalization
3. Learning a Conditional
4. Bayesian Argumentation
5. Conclusions

I. The Main Idea

Deductive Inferences

- Consider the following argument:

P1: It currently rains in Munich.
P2: If it rains, then the streets get wet.

C: Munich's Ludwigstraße is currently wet.

- People familiar with formal logic represent the argument as an instance of *modus ponens*.

$$\frac{\begin{array}{c}A \\ A \rightarrow B\end{array}}{B}$$

- We say that the conclusion follows with necessity, and that we make a mistake if we do not infer B.
- We ask: Are there other (and equally rational) ways to reason with the premisses?

Two Issues

1. The information (i.e. the premisses of the argument) max come from a source which we do not fully trust.
    - We may have listened to the weather forecast and the weather forecast does not say that it will rain.
    - Someone might have told us and we are not sure how reliable this person is.
    - ...
2. Disabling conditions may come to mind.
    - The street might be covered by something which prevents it from becoming wet.
    - there might be strong winds in which case the rain does not have a chance to hit the ground.
    - ...

Upshot

- Taking these concerns into account may lead a rational agent to arrive at a different conclusion.
- We therefore construct a fully Bayesian theory of argumentation which is, or so we hope, in line with how real people reason and which makes nevertheless sense from a normative point of view.
- The theory can be tested and it allows for some flexibility.

The Main Idea – A Skectch

- The agent entertains the propositions A, B,. . .
- The agent represents the causal relations between these propositions in a causal ("Bayesian") network.
- The agent has prior beliefs about the propositions A, B,. . . which are represented by a probability distribution.
- The agent learns new information (i.e. the agent learns the premisses of the argument) and represents them as constraints on the posterior distribution.
- The agent determines the posterior distribution by minimizing some "distance" measure (such as the Kullback Leibler divergence). Idea: we want to change our beliefs in a conservative way.

Learning a Conditional: The General Recipe

If one learns the conditional A $\rightarrow$ B, then

- the new distribution P' has to satisfy the constraint $P'(\mathrm{B}|\mathrm{A}) = 1$
- the causal structure of the problem at hand has to be specified in a causal network.
- the new distribution P' should be as close as possible to the old distribution, satisfying all constraints. One option here is to use the Kullback-Leibler divergence.

If one does this, one can meet a number of challenges (esp. by Douven) and solve van Fraassen's Judy Benjamin Problem.

An Illustration of the Proposed Account: Modus Ponens

- The agent has beliefs about the propositions A and B. These beliefs are represented by a probability function P.
- The agent learns from a perfectly reliable information source that
  - A
  - A $\rightarrow$ B.
- The learned information puts constraints on the new distribution P':
  - A: $P'(\mathrm{A}) = 1$
  - A $\rightarrow$ B: $P'(\mathrm{B}|\mathrm{A}) = 1$.
- The agent then minimizes the KL divergence between P' and P and obtains that $P'(\mathrm{B}) = 1$. This is what we would expect from *modus ponens*.

My Goal: Justify this approach!

II. The Kullback-Leibler Divergence and Conditionalization

The Kullback-Leibler Divergence

- Let $\mathrm{S_1}, \ldots, \mathrm{S_n}$ be the possible values of a random variable S over which probability distributions P and P' are defined.
- The Kullback-Leibler divergence between P' and P is then given by

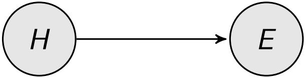
$$D_{KL}(P'||P) := \sum_{i=1}^{n} P'(\mathrm{S_i}) \log \frac{P'(\mathrm{S_i})}{P(\mathrm{S_i})} \, .$$

- Note that the KL divergence is not symmetrical. So it is not a distance.
- Note also that if the old distribution P is the uniform distribution, then minimizing the Kullback-Leibler divergence amounts to maximizing the entropy.
- There are other distance measures and I consider it to be (at least partly) an empirical question which measure is best. Example:

$$D(P'||P) := \sum_{i=1}^{n} (\sqrt{P'(\mathrm{S_i})} - \sqrt{P(\mathrm{S_i})})^2 \, .$$

Conditionalization

- We introduce the binary propositional variables H and E:
  H: "The hypothesis holds", and ¬H: "The hypothesis does not hold".
  E: "The evidence obtains", and ¬E: "The evidence does not obtain".
- The probabilistic relation between H and E can be represented in a Bayesian Network:



- We set $P(\mathrm{H}) = h$ and $P(\mathrm{E}|\mathrm{H}) = p, P(\mathrm{E}|\neg\mathrm{H}) = q$.

Conditionalization

- Calculate the prior distribution over H and E. ($\overline{x} := 1 - x$)

$$P(\mathrm{H},\mathrm{E}) = h\,p \quad , \quad P(\mathrm{H},\neg\mathrm{E}) = h\,\overline{p}$$
$$P(\neg\mathrm{H},\mathrm{E}) = \overline{h}\,q \quad , \quad P(\neg\mathrm{H},\neg\mathrm{E}) = \overline{h}\,\overline{q}\,.$$

- Next, we learn that E obtains, i.e. $P'(\mathrm{E}) = 1$.
- We assume that the network stays the same as before. Hence

$$P'(\mathrm{H},\mathrm{E}) = h'\,p' \quad , \quad P'(\mathrm{H},\neg\mathrm{E}) = h'\,\overline{p'}$$
$$P'(\neg\mathrm{H},\mathrm{E}) = \overline{h'}\,q' \quad , \quad P'(\neg\mathrm{H},\neg\mathrm{E}) = \overline{h'}\,\overline{q'}\,.$$

- From $P'(\mathrm{E}) = h'\,p' + \overline{h'}\,q' = 1$, we conclude that $p' = q' = 1$.
- Minimize the KL divergence: $P'(\mathrm{H}) = P(\mathrm{H}|\mathrm{E})$
- Note that Jeffrey conditionalization obtains if one learns E with $P'(\mathrm{E}) =: e' < 1$.

Learning a Conditional

- Let us now apply this methodology to learning a conditional. We start with a prior probability distribution P over the variables A and B. Next, we learn that $\mathrm{A} \rightarrow \mathrm{B}$, i.e. we impose the constraint $P'(\mathrm{B}|\mathrm{A}) = 1$ on the new distribution P'. Minimizing the KL divergence between P' and P leads to $P'(\mathrm{A}) \leq P(\mathrm{A})$. In fact, one obtains exactly the same result as for conditioning on the material conditional.
- Note also that $P(\mathrm{A}|\mathrm{B}, \mathrm{A} \rightarrow \mathrm{B}) = P(\mathrm{A}|\mathrm{B})$, if A → B is the material conditional. One obtains the same result if one minimizes KL after having learned B and A → B. This observation is related to the *Old Evidence Problem*, which is considered to be a challenge for Bayesianism.

III. Learning a Conditional

The Ski Trip Example

Harry sees his friend Sue buying a skiing outfit. This surprises him a bit, because he did not know of any plans of hers to go on a skiing trip. He knows that she recently had an important exam and thinks it unlikely that she passed. Then he meets Tom, his best friend and also a friend of Sue, who is just on his way to Sue to hear whether she passed the exam, and who tells him,

If Sue passed the exam, then her father will take her on a skiing vacation.

The Ski Trip Example

Harry sees his friend Sue buying a skiing outfit. This surprises him a bit, because he did not know of any plans of hers to go on a skiing trip. He knows that she recently had an important exam and thinks it unlikely that she passed. Then he meets Tom, his best friend and also a friend of Sue, who is just on his way to Sue to hear whether she passed the exam, and who tells him,

If Sue passed the exam, then her father will take her on a skiing vacation.

Recalling his earlier observation, Harry now comes to find it more likely that Sue passed the exam.

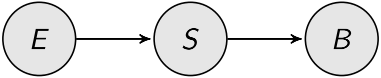
Ref.: Douven and Dietz (2011)

Modeling the Ski Trip Example

We define three variables:

- E: Sue has passed the exam.
- S: Sue is invited to a ski vacation.
- B: Sue buys a ski outfit.

The causal structure is given as follows:



Additionally, we set $P(\mathrm{E}) = e$ and

$$P(\mathrm{S}|\mathrm{E}) = p_1 \quad , \quad P(\mathrm{S}|\neg\mathrm{E}) = q_1$$
$$P(\mathrm{B}|\mathrm{S}) = p_2 \quad , \quad P(\mathrm{B}|\neg\mathrm{S}) = q_2.$$

Note that the story suggests that $p_1 > q_1$ and $p_2 > q_2$.

The Ski Trip Example

- Learning: $P'(\mathrm{B}) = 1$ and $P'(\mathrm{S}|\mathrm{E}) = 1$.
- Again, the causal structure does not change.

  Theorem: *Consider the Bayesian Network above with the prior probability distribution. Let*

$$k_0 := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2}.$$

  We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned information is modeled as constraints on P', and (iii) P' minimizes the Kullback-Leibler divergence to P. Then $P'(\mathrm{E}) > P(\mathrm{E})$, iff $k_0 > 1$.

- The same result obtains for the material conditional.

The Ski Trip Example: Assessing k_0

1. Harry thought that it is unlikely that Sue passed the exam, hence e is small.
2. Harry is surprised that Sue bought a skiing outfit, hence

$$P(\mathrm{B}) = e\,(p_1\,p_2 + \overline{p_1}\,q_2) + \overline{e}\,(q_1\,p_2 + \overline{q_1}\,q_2)$$

   is small.
3. As e is small, we conclude that $q_1\,p_2 + \overline{q_1}\,q_2 := \epsilon$ is small.
4. p_2 is fairly large (≈ 1), because Harry did not know of Sue's plans to go skiing, perhaps he even did not know that she is a skier. And so it is very likely that she has to buy a skiing outfit to go on the skiing trip.
5. At the same time, q_2 will be very small as there is no reason for Harry to expect Sue to buy such an outfit in this case.
6. p_1 may not be very large, but the previous considerations suggest that $p_1 \gg \epsilon$.

The Ski Trip Example: Assessing k_0

We conclude that

$$k_0 := \frac{p_1\, p_2}{q_1\, p_2 + \overline{q_1}\, q_2} = \frac{p_1}{\epsilon} \cdot p_2$$

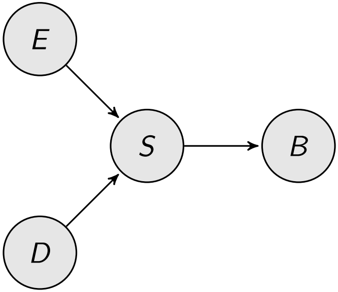
will typically be greater than 1. Hence, $P'(\mathrm{E}) > P(\mathrm{E})$.

No Causal Structure

- What if no causal structure is imposed?
- We computed this case, i.e. we considered only the three variables B, E and S and modeled the learning of $P'(\mathrm{B}) = 1$ and $P'(\mathrm{S}|\mathrm{E}) = 1$ in the usual way.
- Minimizing the KL divergence then leads to $P'(\mathrm{E}) < P(\mathrm{E})$, i.e. to the wrong result.

Disabling Conditions

- A disabling condition D could obtain.
- Then the modified network looks as follows.



The Ski Trip Example Revisited

- Learning: $P'(\mathrm{S}|\mathrm{E}, \neg\mathrm{D}) = 1$ and, as before, $P'(\mathrm{B}) = 1$.
- Then the following theorem holds:

Theorem: *Consider the Bayesian Network in Figure 7 with a prior probability distribution. Let*

$$k_d := \frac{p_1\, p_2}{q_1\, p_2 + (\overline{q_1} - d)\, q_2}.$$

We furthermore assume that (i) the posterior probability distribution P' is defined over the same Bayesian Network, (ii) the learned information is modeled as constraints on P', and (iii) P' minimizes the Kullback-Leibler divergence to P. Then $P'(\mathrm{E}) > P(\mathrm{E})$, iff $k_d > 1$. Moreover, if $k_d > 1$ and $p_2 > q_2$, then $P'(\mathrm{D}) < P(\mathrm{D})$.

Judy Benjamin Problem

A soldier is dropped with her platoon in a territory that is divided in two parts, the Red Territory (R) and the Blue Territory (¬R) where each territory is also divided in two parts, Second Company (S) and Headquarters Company (¬S), forming four sections of almost equal size. The platoon is dropped somewhere in the middle so she finds it equally likely to be in one section as in any of the others, i.e. $P(\mathrm{R}, \mathrm{S}) = \mathrm{P}(\mathrm{R}, \neg\mathrm{S}) = \mathrm{P}(\neg\mathrm{R}, \mathrm{S}) = \mathrm{P}(\neg\mathrm{R}, \neg\mathrm{S}) = 1/4$. Then they receive a radio message:

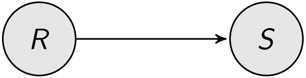
I can not be sure where you are. If you are in Red Territory the odds are 3:1 that you are in the Secondary Company.

How should Judy Benjamin update her belief function based on this communication?

Ref.: van Fraassen (1981)

Judy Benjamin Problem

- We introduce two binary propositional variables. The variable R has the values R: "Judy lands in Red Territory", and ¬R: "Judy lands in Blue Territory". The variable S has the values S: "Judy lands in Second Company", and ¬S: "Judy lands in Headquarters".
- The probabilistic relation between the variables:

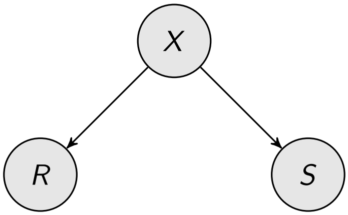


- Learning: $P'(\mathrm{S}|\mathrm{R}) = k \neq 1/2$
- Assume that the network does not change. Then minimizing the Kullback-Leibler divergence yields $P'(\mathrm{R}) < P(\mathrm{R})$, which is not intuitive.

Properly Modeling the Judy Benjamin Problem

We define:

- R: The platoon is dropped in the Red Territory.
- S: The platoon is dropped in the Secondary Company.
- X: Wind comes from a certain direction (or any other cause that comes to mind).



The Judy Benjamin Example

- Learning: $P'(\mathrm{S}|\mathrm{R}) = k \neq 1/2$
- Then the following theorem holds:

Theorem: *Consider the Bayesian Network above with a suitable prior probability distribution P. We furthermore assume that (i) P' is defined over the same Bayesian Network, (ii) the learned information is modeled as a constraint on P', and (iii) P' minimizes the Kullback-Leibler divergence to P. Then $P'(\mathrm{R}) = P(\mathrm{R})$.*

No Causal Structure

- Perhaps it is enough to only take the existence of a third variable X into account, without imposing a causal structure. Let us compute this case!
- We find that imposing $P'(\mathrm{S}|\mathrm{R}) = k \neq 1/2$ as a constraint on the posterior distribution and minimizing the KL divergence leads to $P'(\mathrm{R}) < P(\mathrm{R})$, i.e. to the wrong result.

Why Accepting the Proposed Account

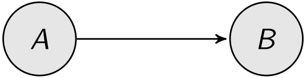
Worry: Our account misses a top-down (axiomatic, Dutch book etc.) justification. However:

1. It is non-trivial that our account provides the right answers in all considered cases.
2. In some cases it also forced us to reconsider our intuitions.
3. There is no other account that achieves this.
4. Using other "distances" (e.g. the Euclidean distance) instead of KL leads to wrong results.
5. Not fixing the correct causal structure (or perhaps better: the proper conditional independencies) leads to wrong results.

Upshot: Taken together, these points provide a strong justification for the proposed account.

IV. Bayesian Argumentation

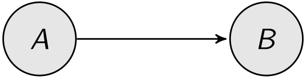
Modus Ponens



$$\frac{\begin{array}{c}\text{A}\\ \text{A} \rightarrow \text{B}\end{array}}{\text{B}}$$

- Prior distribution: $P(\mathrm{A}) = a, P(\mathrm{B}|\mathrm{A}) = p, P(\mathrm{B}|\neg\mathrm{A}) = q$. Hence $P(\mathrm{A},\mathrm{B}) = a\,p,\ P(\mathrm{A},\neg\mathrm{B}) = a\,\overline{p},\ P(\neg\mathrm{A},\mathrm{B}) = \overline{a}\,q,\ P(\neg\mathrm{A},\neg\mathrm{B}) = \overline{a}\,\overline{q}$.
- We learn (i) $P'(\mathrm{A}) = a' = 1$ and (ii) $P'(\mathrm{B}|\mathrm{A}) = p' = 1$. Hence, $P'(\mathrm{A},\mathrm{B}) = 1,\ P(\mathrm{A},\neg\mathrm{B}) = P(\neg\mathrm{A},\mathrm{B}) = P(\neg\mathrm{A},\neg\mathrm{B}) = 0$.
- The constraints uniquely fix the posterior distribution: $P'(\mathrm{B}) = P'(\mathrm{A},\mathrm{B})/P'(\mathrm{A}|\mathrm{B}) = P'(\mathrm{A},\mathrm{B})/P'(\mathrm{A}) = 1$.

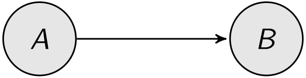
Modus Tollens



$$\frac{\begin{array}{c}\neg\,\mathrm{B} \\ \mathrm{A} \rightarrow \mathrm{B}\end{array}}{\neg A}$$

- Prior distribution: $P(\mathrm{A}) = a, P(\mathrm{B}|\mathrm{A}) = p, P(\mathrm{B}|\neg\mathrm{A}) = q$. Hence $P(\mathrm{A},\mathrm{B}) = a\,p$, $P(\mathrm{A},\neg\mathrm{B}) = a\,\overline{p}$, $P(\neg\mathrm{A},\mathrm{B}) = \overline{a}\,q$, $P(\neg\mathrm{A},\neg\mathrm{B}) = \overline{a}\,\overline{q}$
- We learn (i) $P'(\mathrm{B}) = 0$ and (ii) $P'(\mathrm{B}|\mathrm{A}) = 1$. Hence, $a' + \overline{a'}\,q' = 0$, hence $a' = q' = 0$. Hence, $P'(\mathrm{A},\mathrm{B}) = P(\mathrm{A},\neg\mathrm{B}) = P(\neg\mathrm{A},\mathrm{B}) = 0$, $P(\neg\mathrm{A},\neg\mathrm{B}) = 1$.
- The constraints uniquely fix the posterior distribution: $P'(\mathrm{A}) = 0$.

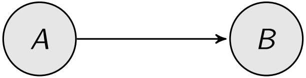
Affirming the Consequent



$$\begin{array}{c} \text{B} \\ \text{A} \rightarrow \text{B} \\ \hline \text{A} \end{array}$$

- Prior distribution: $P(\mathrm{A}) = a, P(\mathrm{B}|\mathrm{A}) = p, P(\mathrm{B}|\neg\mathrm{A}) = q$. Hence $P(\mathrm{A},\mathrm{B}) = a\,p$, $P(\mathrm{A},\neg\mathrm{B}) = a\,\overline{p}$, $P(\neg\mathrm{A},\mathrm{B}) = \overline{a}\,q$, $P(\neg\mathrm{A},\neg\mathrm{B}) = \overline{a}\,\overline{q}$
- We learn (i) $P'(\mathrm{B}) = 1$ and (ii) $P'(\mathrm{B}|\mathrm{A}) = 1$. Hence, $p' = q' = 1$.
- Minimizing the Kullback Leibler divergence yields $P'(\mathrm{A}) = P(\mathrm{A}|\mathrm{B})$.

Denying the Antecendent

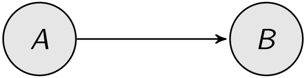


$$\neg\, \text{A}$$
$$\text{A} \rightarrow \text{B}$$

$$\neg B$$

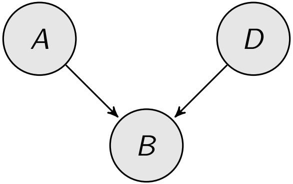
- Prior distribution: $P(\text{A}) = a, P(\text{B}|\text{A}) = p, P(\text{B}|\neg\text{A}) = q$. Hence $P(\text{A}, \text{B}) = a\, p$, $P(\text{A}, \neg\text{B}) = a\, \overline{p}$, $P(\neg\text{A}, \text{B}) = \overline{a}\, q$, $P(\neg\text{A}, \neg\text{B}) = \overline{a}\, \overline{q}$
- We learn (i) $P'(\text{A}) = 0$ and (ii) $P'(\text{B}|\text{A}) = 1$. Hence, $a' = 0$ and $q' = 0$.
- Minimizing the Kullback Leibler divergence yields $P'(\text{B}) = P(\text{B}|\neg\text{A})$ and hence $P'(\neg\text{B}) = P(\neg\text{B}|\neg\text{A})$.

Relaxing Idealizations I: Doubting a Premisse



- For example, in the case of *modus ponens* we assume that
  - $P'(\mathrm{A}) < 1$, and
  - $P'(\mathrm{B}|\mathrm{A}) = 1$.
- Minimizing the Kullback Leibler divergence, one then obtains that $P'(\mathrm{B}) < 1$.

Relaxing Idealizations II: Disabling Conditions



- D is a disabling condition, such as the presence of wind, that the street is covered,. . .
- Here we have $P(\mathrm{D}) > 0$.
- In the case of *modus ponens*, we learn that
  - $P'(\mathrm{A}) = 1$
  - $P'(\mathrm{B}|\mathrm{A}, \neg\mathrm{D}) = 1$.
- Minimizing the Kullback Leibler divergence yields $P'(\mathrm{B}) < 1$.

V. Conclusions

Conclusions

1. We have sketched a unified Bayesian account of argumentation.
2. Getting the causal structure right turned out to be significant.
3. We depart from mainstream Bayesianism and include different distance measures which can be empirically investigated.
4. The proposed account is normatively interesting and it can be tested empirically in a systematic way.

There is much more to do...

Thanks. . .

. . . for your attention!