Marcello's Comments on the Chapter about Weight

9/1/2022

1 Purpose of this document

Marcello's comments on the chapter about weight.

2 Weight and completeness (Sec 5)

Rafal writes:

> So the second difficulty is that on this approach the weight of evidence becomes very sensitive not only to what the actual evidence is, but also to what an ideal evidence in a given case should be. unless as clear and epistemologically principled guidance as to how to formulate such ideal lists is available, this seems to open a gate to arbitrariness. Change of awareness of one's own ignorance, without any major chance to the actual evidence obtained, might lead to overconfidence or under-confidence in one's judgment. Moreover, it is not clear how disagreement about weight arising between agents not due to evidential differences, but rather due to differences in their list of ideal items of evidence should be adjudicated.

I am not sure. What makes a body of evidence complete is subjective, since it depends on what one knows about a given situation. But this fact cannot be used as a criticism for a theory of weight based on completeness. If I did not know the defendant had a huge archive of documents in his office, I might think that all the evidence I have (without the documents in the archive) is complete. But when I learn about the existence of that archive and realize that the evidence lacks the documents in the archive, then my evidence is clearly incomplete. This does not makes the assessment of weight-as-completeness subjective. This is just how things should be.

Perhaps there are different levels of analysis here:

- assess completeness of evidence based on an ideal list that would apply universally to any case like the one under consideration (script approach)
- assess completeness based on a specific recounting of what happened that is agreed by both parties (shared narrative approach).
- assess completeness based on a specific recounting of what happened pout forward by one of the parties (partisan narrative approach).

Arbitrariness might exist in the script approach (we might disagree about the right script to apply to determine the ideal list of items of evidence), but does not exist in the narrative approach.

3 Imprecise probabilities (Sec 8 and 9)

These two sections are very interesting, but we need to think about they can fit in the chapter as such. Most of the examples (and counterexamples) in these sections are about coin tossing and sample size. I think we might need to consider examples of quantitative evidence in the law, DNA evidence, multiple reference populations or different sample sizes. To warrant a discussion of imprecise probabilities here, we need to show that imprecise probability measures (and also Joyce's notion of weight) have a prima facie applicability to the law and then we can show that they are inadequate for various reasons.

The bit about proper scoring rules and teh Brier score is particularly interesting. This can perhaps belong to the section on accuracy. But it does seem a requirement of any weight measure, juyst like any probability measures, that we can connect it up to accuracy in some way.

4 Sec 10 (highere order approach)

The general idea here is clear, but I wonder if a simple example or two for legal application is helpful. Perhaps an example with DNA evidence and sample size or something to that effect.

5 Weight of a distribution (Section 11)

- This and the next one are the crucial sections. This is the map I have now in my mind to follow what is going on in this section:
  i. Notion of information/entropy in general
  ii. Entropy of distributions
  iii. Difference of entropy between distributions (cross-entropy, KL divergence)
  iv. weight of a distribution D is the difference (in some sense to be specificed) between the entropy of the distribution D compared to the uniform distribution (which by default has maximal entropy). But, crucially, this measure iof weight is NOT the KL divergence.
  v. weight of evidence (in the next section)
- The move from the example with the three forks to a distribution of parameter values (each associated with a different probability) is not completely clear.Basically this is the step from (i) "entropy in general" to (ii) entropy of distribution. I can sort of see the connection, but it is not spelled out clearly. This is the part that says "A measure of (lack of) information contained in a whole distribution, is entropy, which is the average Shannon information:..."
- Again, in the discussion of entropy it would be good to have a clear running example, possibly legal in nature.
- Using the grip approximation for continuous distributions is fine, but what is the reason? You seem to say it is because we will compare continuous and discrete distribution. That seems sensible. Can you give an example?
- KL divergence. This is the difference between the entropy associated between the two distribution and because of the properties of logarithm it is the log of the ratio. Right?
- After that, you write "The idea is that the more informative a piece of evidence is, as compared to the uniform distribution, the more weight it has, on scale 0 to 1:" Are you talking about a piece of *evidence* or a *distribution*?
- Footnote 14 seems crucial here ("The reader might ask: why not to use the Kullback-Leibler divergence from the uniform distribution instead? Because this divergence does not measure the difference in how informed the distribution is. "). This might need more explanation. But if the KL divergence does not measure how informative a distribution is, then what does it measure? What are the sues for KL measure as opposed to the weight measure you propose?Also, I think one can understand your measure of weight without the KL divergence measure, which seems to be more complicated and meant to do something else.
- Another question is why the weight of a distribution P isn't simply 1-H(P). What are the advantages of comparing with the uniform distribution and using 1-(H(P)/H(uniform)? Another option for weight of a distribution would be the difference between the entropy of the uniform and the entropy of P, that is, H(uniforms)-H(P). Is this equivalent to something like the log of the ratio of the distribution, log(P/uniform)? Why then not take 1-log(P/uniform)? At any rate, more explanation why you picked that measure and excluding others in the vicinity would be useful.
- The graphs used to compare Urbaniak-weight with Joyce-weight need additional explanation. You earlier complained that Joyce-weight display strange patterns. But O noticed that the graph of Urbaniak-weight are similar to those of Joyce-weight. So do you see strange patterns here too, or not? For example, why does the weight of a beta distribution drop as the ratio heads/tails become less sweked, but suddenly increases when teh ratio is close to 50/50? Is this a normal behavior? You might to spend mroe time, suggesting that this behavior is intuitive and to be expected.

- You say: "the entropy decreases with the number of observations". A beta(50,50) will have greater weight than beta(5, 5). I can see why that is because beta(50,50) is more sharply concentrated around .5 than beta(5,5). But some explanatio might be helpful.
- You also say: "it [=entropy] decreases faster if the proportions are closer to the extremes. " Beta(80,20) should presumably have more weight than beta(50,50) since it is more extreme. Why is that exactly? Why is it that a more skewed distribution is weightier? I can sort of see why, but more explanation would be needed. The problem here is that—looking the general graph of beta distribution and also from the comparison with Joyce— it seems that beta(50,50) is weightier than beta (55,45). There is slight increase in weight when we get to 50/50. Why is that? is this intended? Isn't that alike one of the strange patterns that we see in Joyce-weight?
- There is a potential ambiguity (or potential source of confusion) in the illustration using beta distributions. The guiding intuitions is that beta(a, b) represents a set of observations with a successes and b failures. So the beta distribution in fact represent a bunch of observations. Under this interpretation, beta(50,50) is weightier than beta (5,5) because it represents more observations. So, in some sense, here we are not thinking about the weight of a distribution, but the weight of the observations (the more observation, the weightier they are). At the same time, your account shows that the mere number of observation does not solely determine weight. What also matters is how the outcomes of he observation are allocated (the proportion of successes versus failures). So, then, it seems that weight tracks (a) how many observations there are and (b) how they are distributed of structured? So, the question is, when we talk about the weight of a distribution, are we talking about anything other than the weight of (a) and (b)—i.e. observation plus how they are structured?
- One interesting thing to know would be, when do two beta distributions (with different levels of proportion) have the same weight? For example, consider beta(80,20), so the proportion here is 80/20. Now let's take a proportion like 60/40? What would be a beta distribution that satisfies this proportion and that has the same weight as beta(80,20)? I suppose that the beta distribution will have to have more sample observations. But how many more?
- Here is another way of thinking about this problem intuitively. Suppose I want to know whether this coin is fair (0.5). Contrast this with suppose I want to whether the coin is biased with bias 0.8 heads. The question is how many observations, and in what proportions, would I need to be equally sure (=to have equally weighty observations) to conclude that the coin is fair versus 0.8 biased toward heads?

6 Weight of evidence (Sec 12)

- The start of this section is revealing:

  > "So far we have discussed the weight of a distribution, meant to measure how informed an agent is about an issue. If the agent starts with a uniform prior, this is a good enough approximation of how informed the evidence made them. But in general, how much more information is obtained is context-dependent. We want a prior-relative notion of weight, following the intuition that weight consideration should guide our information gathering also in making us stop collecting further evidence in light of what we already know. But for weight of evidence to have this feature, it has to depend on what we already know."

  So, if I understand this right, when you talk about the "weight of a distribution", you are actually talking about the "weight of the evidence/observations" assuming you started out from uniform prior. This makes sense since the weight of a distribution is measured against the baseline of a uniform distribution. Ss my interpretation right?
- Earlier I thought that "weight of a distribution" and "weight of evidence" were two different things. Now, if I understand this right, looks like "weight of evidence" is a generalization of the idea of weight of a distribution. Weight of a distribution is the weight of the evidence when the prior is uniform. Is this right?
- If the preceding point is right, it might also address the ambiguity I pointed out in my earlier remarks–ambiguity between "weight of distribution" and "weight of the observations" (which underline the distribution).

- The way you suggest to measure the weight of the evidence is as follows:

  > In a given context, consider your distribution for the target hypothesis H given what you already know. Then update on the evidence. This might increase the weight for H, if the evidence confirms your conviction, or decrease it, if it goes against what the previous evidence tells you. Take the difference between the prior weight and the posterior weight (Δw) as your measure of the weight of evidence in that context.

  The idea is to take the difference between W(prior distribution) and W(posterior distribution). But, in light my remarks above, I wonder why "weight of evidence" could be measured, follows:

$$W(E) = 1 - \frac{H(\textit{posterior distribution (after updating on E)}}{H(\textit{prior distribution})}$$

  This is just a generalization of your earlier measure of weight of distribution. Your measure of the weight of a distribution had in the denominator the uniform distribution. Now, instead of the uniform, the weight of evidence has the prior distribution (which can be any distribution)

$$W(P) = 1 - \frac{H(\textit{distribution P}}{H(\textit{uniform distribution})}$$

  Any thought about this approach and your approach in terms of delta-weight? Are the two approaches equivalent?
- The discussion of the rocking and the abused child is interesting, but not completely clear yet.

7 Weight and Accuracy (Sec 13)

- A general thought about weight and accuracy is this: a probability assessment that is based on more evidence (or weightier evidence) will tend to be more accurate (in the sense of being closer to the true value) than a probability estimate based on less evidence (less weighty evidence). Hamer show that more evidence reduced uncertainty (probability assessments will tend to be more concentrated toward 0 and 1, assuming true values are 0 and 1).
- If I understand the strategy correctly, the idea is this. (a) The first step is to formulate something like the Brier score but for distribution. You suggest to use the KL measure between the indicator distribution (true distribution) and the distribution in question. This seems right as intuitively it measure how distant a distribution is from the true distribution. (b) The second step is to show that a weightier distribution will also have a better KL measure. Is this right? The question, how does this map onto the general idea that weightier evidence improves accuracy?