Notes on Uncertainty

Wilker Aziz

September 29, 2020

In statistics *uncertainty* refers to a state of limited knowledge about a variable of interest. Variables which we are uncertain about are also called random variables. The probability distribution of a random variable (rv) is a representation of our uncertainty about the outcomes of the rv.[1] Uncertainty quantification may take different forms in different disciplines, but is best understood under the lens of probability theory.

In Bayesian statistics uncertainty is quantified (or rather, revised) by direct application of probability calculus (in particular, Bayes rule). Given a model of the interaction amongst the random variables of interest, Bayes rule specifies how one should update their beliefs about unobserved random variables (or estimands) in light of observations.[2] If the estimand is a hypothesis θ, we are talking about the posterior distribution $\Theta|\alpha, \mathcal{D}$, where $\Theta|\alpha$ is a prior distribution over hypotheses and $\mathcal{D}$ is a collection of observations.[3] Uncertainty quantification in Bayesian statistics takes only acknowledging that our estimands are random, probability calculus does the rest.[4] Making inferences about parameters is one scenario of interest. Making predictions for novel input is another. Luckily the Bayesian statistician does not have to invent any new machinery, instead, she uses the posterior predictive distribution $Y * |x_*, \alpha, \mathcal{D}$, which again, given a model, follows by probability calculus.[5]

Non-Bayesian views. You will find mentions to 'uncertainty quantification' all over the literature, yet I struggle identifying a proper characterisation of the term, especially so when we are interested in quantifying uncertainty about

[1] Mind the confusion. Uncertainty is not a number. For example, the variance of an rv is not our uncertainty about that rv. Variance is a functional of the rv, that is, a numerical assessment of a function of the rv (in this case $(X - \mathbb{E}[X])^2$, for an rv X) computed in expectation w.r.t. the rv's distribution.

[2] Recall that Bayes rule gets a special name, but it's a trivial consequence of the axioms of probability theory.

[3] Initially, the prior distribution captures our uncertainty about Θ, after we observe data, the posterior distribution captures our revised uncertainty.

[4] Well, to be more precise, it takes specifying a full model (including conditional independence assumptions and parameterisation) of the rvs in consideration.

[5] Do you see why it's so difficult to argue against Bayes? All we ever do is to use probability calculus over and over :-)

unobserved rvs. I'm not particularly satisfied with the Frequentist view, which usually involves hypothetical repetitions of an experiment: I rarely can afford repeating data collection, for example, and even if I can, it's hardly ever the case that I can control for all variables as to guarantee that repetitions are 'exchangeable' in some sense (if you note the vagueness I have to resort to, it's not my fault, it's because there is no 'theory' in this case, but rather a collection of principles that are invoked in specific cases). The best I gather is that in machine learning, we will take as uncertainty quantification any process that leads to an early estimate of our chances of being wrong in the future (again, pretty vague). For lack of alternative, when working with non-Bayesian views, we will attempt to frame them as approximations to the Bayesian view.

Uncertainty estimates. We are often interested in computing so called 'uncertainty estimates'. Let us concentrate on uncertainty about predictions, which is captured/represented by the posterior predictive distribution. These uncertainty estimates are numerical functions of the rv of interest and its distribution.[6] For example, posterior predictive variance (or Monte Carlo estimates thereof) is one such quantity of interest, the posterior predictive probability of an outcome is another, the expected value of a scoring function is yet another. When working with non-Bayesian models we will substitute the posterior distribution or the posterior predictive distribution by some object defined in a similar domain (for example, we might use a set of stochastic approximate MLE solutions as a substitute for posterior samples, or the likelihood as a *very crude* substitute for the posterior predictive distribution). Generally, we will not be able to afford representing the posterior distribution nor the posterior predictive distribution exactly, instead we will work with samples from processes that approximate those distributions.

Literature

See Gelman et al. (Chapter 1 and Chapter 4; 2013).

References

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition.

[6] For brevity, you could say *a functional of the rv*.