Statistical Decision Theory and Bayesian Analysis

Chapter 3: Prior Information and Subjective Probability

Lili MOU

moull12@sei.pku.edu.cn
http://sei.pku.edu.cn/~moull12

11 May*April* 2015

Reference

§3, James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, 1985.

This book covers basic materials of statistical decision theory in an easy-to-understand yet critical manner. The prerequisite is rather low.

- Statistical level: moderately serious statistics
- Mathematical level: easy advanced calculus

This slide mainly picks textual materials in Chapter 3. For detailed math, please refer to other resources.

Subjective Probability

Some even argue that the frequency concept never applies, it being impossible to have an infinite sequence of i.i.d. repetitions of any situation, except in a certain imaginary (subjective) sense.

Criticisms on Noninformative Priors

- Depend on experimental structure
- "Marginalization paradox"
- *Perhaps the most embarrassing feature of noninformative priors, however, is simply that there are often so many of them.*

Hierarchical Priors

- It is somewhat difficult to subjectively specify second stage priors, such sa $\pi_2(\lambda)$ in the above example, but it will be seen in Chapter 4 there is "more robustness" in second stage specifications than in single specifications (i.e., there is less danger that misspecification of π_2 will lead to a bad answer).
- The difficulty of specifying second stage priors has made common the use of noninformative priors at the second stage.
- Note that a hierarchical structure is merely a convenient representation for a prior, rather than an entirely new entity; any hierarchical prior can be written as a standard prior.

$$\pi(\boldsymbol{\theta}) = \int_{\Lambda} \pi_1(\boldsymbol{\theta}|\lambda)\, \mathrm{d}F^{\pi_2}(\lambda)$$

Criticisms

Few statisticians would object to a Bayesian analysis when θ was indeed a random quantity with a known prior distribution, or even when the prior distribution could be estimated with reasonable accuracy (from, say, previous data). The major objections are to use of subjective or "formal" prior distributions, especially when θ is random only in a subjective sense.

Objectivity

To most non-Bayesians, classical statistics is "objective" and hence suitable for the needs of science, while Bayesian statistics is "subjective" and hence (at best) only sueful for making personal decisions. Bayesians respond to this in several ways.
. . . very few statistical analyses are even approximately "objective."

It is indeed rather peculiar that some decision-theorists are ardent anti-Bayesians, precisely because of the subjectivity of the prior, and yet have no qualms about the subjectivity of the loss.

Model assumptions are prior *per se*

Box (1980) says

> *In the past, the need for probabilities expressing prior belief has often been though of , not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief. . . I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters.*

More bluntly, Good (1973) says

> *The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science.*

Data Snooping

> *The potential for misuse by careless or unscrupulous people is obvious.*

Savage (1962):

> *It takes a lot of self-discipline not to exaggerate the probabilities you would have attached to hypotheses before they were suggested to you*

Digression on the Machine Learning Research Regime

Data snooping is inevitable, at least in machine learning research.

- Suppose we have proposed a new algorithm, or model whatever.
- The **only right** regime is to train the model on the training set, validate it on the CV set, until satisfied, test it on the test set **only once**, and report it in the paper.
- A severe problem rises if your test accuracy does not outperform the state-of-the-art result, you really have no chance to publish your work, never ever.
- If you then improve your model, you constituent **data snooping**.

Digression on the Machine Learning Research Regime (2)

Non-monotonicity is a peculiar property of machine learning research.

- If your proposed algorithm yields 61% CV accuracy and 60% test accuracy, you are happy because the state-of-the-art result is 59%, say.
- You hope to further tune hyperparameters to persue a better result. Unfortunately, you achieve 64% CV acc. but 58% test acc.
- Now comes the problem. You have to dwarf the accuracy to 58% in your paper (which would not be actually accepted).
- Some researchers may change the model a little bit, so that a new approach comes into birth, yielding 61% accuracy for both validation and testing. It is a relief to the researchers, because they are not cheating (at least ostensibly).
- Indeed, they are, because (1) the test set can be used at most once, and (2) Box (1980) suggests no logical difference between models, parameters, or here hyperparameters.
- Ironically, the above unfavorable hyperparameter only affects your research career. Other researchers do not care about it.
- The "training–validating–testing" regime should be applied to all researchers, because CV acc is an unbiased estimation of test acc. All of us anticipate a high test acc, and thus should have choose the highest cv acc in the world. However, even if you announce your results, other researchers pretend not to have seen it. ☹