Research Proposal

Bayesian foundations and loss function based inference

József Konczer
konczer.j@gmail.com

Introduction

"Probability is the very guide of life", however, there are many ways to interpret the concept, and its guidance.

The Bayesian interpretation suffers from the lack of justification of prior choices. There have been many attempts to find appropriate principles for Objective Bayesianism [1, 2], besides the construction of a Subjective Bayesian alternative pioneered by Savage and de Finetty.

Von Neumann's minimax theorem and the framework of game theory served another possible justification of inference methods. The 1950 book of Abraham Wald [3] defines inference as an optimal strategy followed by "Experimenter" who plays a zero sum game against "Nature". In this early work only a quadratic loss function was analysed, however, independently, R. L. Kashyap investigated zero sum games with relative entropy loss function in the 70's [4, 5].

Unfortunately, the game theoretic interpretation did not gain popularity in statistics, but similar concepts appeared in decision theory called Maxmin Expected Utility (MEU) [6] and in Machine Learning known as Generative Adversarial Network (GAN) [7].

My main aim is to revitalize and further expand the concept of a loss function based inference, which could treat many different areas, such as decision theory, statistical inference, data compression, statistical physics and quantum information theory in one coherent framework.

If we extend the framework to more interactive data acquisition procedures, then it can give a non circular definition for the experimenter's subjective probability concept, who gathered some data, have a model and objectives, but faces an uncertain situation.

Bayesian methods are becoming increasingly popular in Machine Learning and engineering applications due to sufficient computational power. However, the choice of prior is often not well justified, which can cause problems mainly in data scarce situations. A loss function based approach can increase accuracy on small data sets. It would also improve the accuracy of prior dominated models, which are typically the ones having many parameters.

Formal definition of the model

To define an inference game, one first needs a probabilistic model, which contains a data domain $\mathcal{D}$, a target set $\mathcal{T}$, a parameter set Θ (for the sake of simplicity these domains can be defined as finite dimensional manifolds), a parameter dependent PDF function on $\mathcal{D}$, $f(.|\theta)$ and a parameter dependent PDF function on $\mathcal{T}$, $g(.|\theta)$ (see Figure 1). As an action, for any data point $x \in \mathcal{D}$, the experimenter can construct a data dependent predictive PDF on $\mathcal{T}$, $h(.|x)$. (This move represents the inference made by the experimenter.) The game's payoff function will be a functional $L(.||.)$ from a pair of PDF-s on $\mathcal{T}$ to $\mathbb{R}$, and the experimenter's loss will be $L(g(.|\theta)||h(.|x))$ if the "real" distribution in the target space is $g(.|\theta)$, while we predicted an $h(.|x)$ distribution based on our collected data $x \in \mathcal{D}$. The three spaces and two functions define a statistical model $\mathcal{M} = \{\mathcal{D}, \mathcal{T}, \Theta, f, g\}$ and a statistical model together with a loss functional define an inference game $\mathcal{G} = \{\mathcal{M}, L\}$. In this game "Nature" can use a mixed strategy π to choose from the parameter set Θ i.e. π is a distribution on Θ, and the experimenter can pick any predictive function h to minimize its expected loss, defined by:

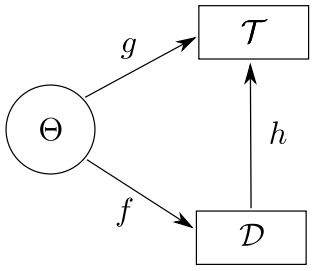

Figure 1: Probabilistic model $\mathcal{M}$

$$\Delta(\pi, h) = \int \mathrm{d}\theta \; \pi(\theta) \, \mathbb{E}_{\xi \sim f(.|\theta)} \left[L(g(.|\theta)||h(.|\xi)) \right] \tag{1}$$

(If the loss functional is convex in the second variable, then the optimal strategies for the experimenter are pure.)

An especially important case is when the loss functional equals to the Kullback–Leibler divergence of "real" and predicted distributions $L(g(.|\theta)||h(.|\xi)) = D_{KL}(g(.|\theta)||h(.|\xi))$. If we choose this special loss functional, then the optimal predictive function will be h^*_π precisely the distribution which would appear after Bayesian inference, taking π as a prior on the parameter space. Furthermore, according to the worst case principle we expect that "Nature" maximizes our loss, using a mixed strategy π^*, even after we use the optimal prediction distribution $h^*_{\pi^*}$. This π^* prior can be viewed as an equilibrium strategy for a zero sum game played between "Nature" and the experimenter. (In case of relative entropy loss functional, the loss can be interpreted as wasted bits between the optimal protocol constructed using $g(.|\theta)$ and a suboptimal protocol derived from $h(.|x)$.)

There are two natural choices for the target space: the data centered (or Aristotelian) and the model centered (or Platonic) approach. If we investigate a probabilistic model, where n i.i.d. variables can be observed, given a parameter dependent PDF on Ω, $f_0(.|\theta)$, a data centered agent would be interested in the most accurate prediction of the distribution of the observable data in the future. After n measurements $\mathcal{M} = \{\mathcal{D} = \Omega^n, \mathcal{T} = \Omega, \Theta, f = f_0^n, g = f_0\}$. On the other hand, a model centered agent would be interested in the inference of model parameters after the measurements $\mathcal{M} = \{\mathcal{D} = \Omega^n, \mathcal{T} = \Theta, \Theta, f = f_0^n, g = \delta\}$ [1].

[1] In the continuous case, one needs to specify the natural coordinates, in which the Dirac-delta function is understood. A natural (and also reparametrization invariant) choice is the Normal coordinates respect to the Fisher information metric.

Already analyzed toy models

I have already started to explore the inference schemes, which can be derived using the inferential game framework. I investigated the simplest possible toy models, where exact calculations and proofs can be made.

I calculated the optimal inferences using relative entropy loss functional in case of finite hypothesis testing problems on binary strings, Bernoulli model, Gauss distributions with unknown mean and variance, and the continuous taxi-cab problem.

In case of binary strings, using an appropriate loss function I was able to derive two very similar versions of Solomonoff's inductive inference.

I generalized the framework to optimal experimental design problems (where the experimenter can choose between possible measurements), and also to the quantum case. The simplest nontrivial case of hypothesis testing of a one qbit system with two alternatives has also been defined and numerically solved.

Proposed work

I would like to continue to gradually explore more complex toy models, e.g., the higher dimensional versions of classical estimation problems, analyzing Markov Chains, Gaussian Processes, Bayesian Linear/Polynomial Regression, and Bayesian Neural Network.

My plans are the following: (1) Explore the effect of the loss functional on the equilibrium inference scheme, and construct an axiomatic framework for loss functionals, suitable for scientific purposes. (2) Understanding frequentist methods as strategies for scoring loss functions. (3) Extend the domains to functional spaces and investigate how the framework could generate empirical distribution functions from data coming from (an unknown) continuous distribution. (4) Use the framework to construct optimal coding protocols (having a probabilistic model of the data, containing uncertain priors), and use these protocols together with the minimum description length (MDL) principle to define and solve model selection problems. (5) Give a non circular interpretation of subjective probability of agents using game theoretical situations. (6) Give an evolutionary interpretation of Bayesian reasoning based on multiplicative betting opportunities using Kelly criterion. (7) Define the problem of Inductive Logic, described by Keynes and Carnap, as a search for the shortest axiom system consistent with true theorems in a formal language, put it into a game theoretic framework, and explore the features of inference methods, optimal for this game. (8) Interpret the methods of statistical physics as optimal strategies in inferential problems. (9) Analyse the framework, when the experimenter can choose between possible experiments, exploring adaptive optimal strategies and heuristics. (10) Generalize and explore the framework suitable for quantum information theory. (11) Investigate appropriate heuristics, iterative methods and approximative inference schemes suitable for complicated, non exactly solvable models. (12) Generalize the concept for Reinforcement Learning, by assuming that the choice of the reward function was made by "Nature" as well, investigate simple toy models, e.g., Two-armed Bandit, and explore how the choice of objectives changes the agent's behaviour.

During the PhD programme, I am planning to engage in multidisciplinary collaborations i.e. with mathematicians, statisticians, economists, computer scientists and physicists, to explore the topic from many sides, and strengthen the coherence between methods used in these fields.

Possible applications

The game theoretic approach, besides serving as a philosophical basis for inference and treating many different procedures in one coherent framework, can also provide effective methods for making predictions without overfitting, and criteria for model selection problems.

The framework can be generalized to other areas, where Bayesian approach is not straightforwardly applicable, e.g., quantum information theory. Furthermore, it can help to assist decisions from gathered data, where different normative considerations can be relevant e.g., in politics, economics and social sciences.

References

[1] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. [Online]. Available: http://www.med.mcgill.ca/epidemiology/hanley/bios601/GaussianModel/JaynesProbabilityTheory.pdf

[2] A. F. M. S. José M. Bernardo, *Bayesian Theory*, 3rd ed., ser. Wiley Series in Probability and Statistics. John Wiley Sons, 2000.

[3] A. Wald., *Statistical decision functions*, ser. Wiley publications in statistics. New York, Wiley, 1950.

[4] R. Kashyap, "Prior probability and uncertainty," *IEEE Transactions on Information Theory*, vol. 17, no. 6, pp. 641–650, Nov. 1971. [Online]. Available: https://doi.org/10.1109/tit.1971.1054725

[5] ——, "Minimax estimation with divergence loss function," *Information Sciences*, vol. 7, pp. 341–364, Jan. 1974. [Online]. Available: https://doi.org/10.1016/0020-0255(74)90021-8

[6] I. Gilboa and D. Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, vol. 18, no. 2, pp. 141–153, Jan. 1989. [Online]. Available: https://doi.org/10.1016/0304-4068(89)90018-9

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf