Meta-Induction, Probability Aggregation, and Optimal Scoring[*]

Christian J. Feldbacher-Escamilla   Gerhard Schurz

Summer 2020

Abstract

[355] In this paper, we combine the theory of probability aggregation with the theory of meta-induction and show that this allows for optimal predictions under expert advice. The full paper to this contribution is published as (Feldbacher-Escamilla and Schurz 2020b, http://doi.org/10.1007/s10472-019-09648-4).

Keywords: meta-induction, probability aggregation, Brier score

1 Introduction

Probability aggregation is an expansion of the theory of judgment aggregation and addresses the question of how to aggregate probability distributions. In past, research in this field centred around the disciplines of economics and political science, law, and philosophy (List and Pettit 2002). Recently, however, increasing work stems also from computer science and artificial intelligence (Rossi, Venable, and Walsh 2011).

We suggest to interpret the weights in characterisation results of linear probability aggregation (cf. section 2) in a success-based way. By cashing out results on no-regret methods for prediction under expert advice of the field of online machine learning (cf. section 3) we show that fixing the parameters in a success-based way allows for optimal probability aggregation (cf. section 4).

[*][This text is published under the following bibliographical data: Feldbacher-Escamilla, Christian J. and Schurz, Gerhard (2020a). "Meta-Induction, Probability Aggregation, and Optimal Scoring". In: *KI 2020: Advances in Artificial Intelligence. Lecture Notes in Computer Science and Artificial Intelligence. Volume 12325*. Ed. by Schmid, Ute, Klügl, Franziska, and Wolter, Diedrich. Cham: Springer Nature, pp. 355–357. DOI: 10.1007/978-3-030-58285-2. All page numbers of the published text are in square brackets. For more information about the underlying project, please have a look at http://cjf.escamilla.academia.name.]

2 Linear Probability Aggregation

The theory of probability aggregation deals with the problem of how to aggregate a set of probability distributions. Abstractly speaking, the question is how to characterise a probability aggregation rule f which takes as input a set of n probability distributions $P_1, \ldots, P_n$ and generates as output a/*the* aggregated probability distribution $P_{aggr} = f(P_1, \ldots, P_n)$. So-called "linear probability aggregation rules" have the following form of a weighted arithmetic mean:

$$P_{aggr} = \sum_{i=1}^{n} w_i \cdot P_i \text{ where } w_i \geq 0 \text{ and } w_1 + \cdots + w_n = 1 \tag{AM}$$

Different interpretations of the weights allow for different specifications. Here we want to argue for interpreting the weights in a regret-based way, because such an interpretation allows for optimal probability aggregation. [356]

3 Optimality in an Expert Advice Setting

In online machine learning, regret bounds of methods for predictions under expert advice are studied (Cesa-Bianchi and Lugosi 2006). The idea is to consider a series of events (E) whose outcomes ($val_t(E)$) have to be predicted by so-called *experts* or *candidate* methods ($P_{1,t}, \ldots, P_{n,t}$ of n candidate methods). Given these predictions, the task is to construct a meta-inductive prediction method $P_{mi,t}$ that uses the candidate method's forecast as input and aims at optimality by approaching the predictive success of the best expert (Schurz 2008, 2019).

We assume that all the mentioned values are within the unit interval. Then we define $P_{mi,t}$ by keeping track of the success rate s of a candidate method i via summing up its score (which is 1 minus the loss l of i's prediction—l is within $[0,1]$ and convex) up to round t and then take the average. Afterwards, we define weights w via cutting off and normalisation (Schurz 2008, sect.1 and sect.7):

$$s_{i,t} = \frac{\sum_{u=1}^{t} 1 - l(P_{i,u}, val_u(E))}{t} \qquad w_{i,t} = \frac{max(0, s_{i,t} - s_{mi,t})}{\sum_{j=1}^{n} max(0, s_{j,t} - s_{mi,t})}$$

If P_{mi} outperforms all other methods, averaging applies, so the weights are always positive and sum up to 1. Based on this, we can define a weighted-average meta-inductive method (MI) as a linear combination (Cesa-Bianchi and Lugosi 2006; Schurz 2008, sect.2.1 resp. sect.7):

$$P_{mi,t+1} = \sum_{i=1}^{n} w_{i,t} \cdot P_{i,t+1} \tag{MI}$$

Regarding the success rate of (MI) one can prove the following bound with respect to the success rates of the candidate methods (Cesa-Bianchi and Lugosi 2006; Schurz 2008, sect.2.1f resp. 7):

Theorem 1. *Given the loss function l is convex it holds:*

$$s_{i,t} - s_{mi,t} \leq \sqrt{n/t} \ \ \forall i \in \{1, \dots, n\}, \text{ so } \lim_{t\to\infty} max(s_{1,t}, \dots, s_{n,t}) - s_{mi,t} \leq 0$$

This theorem shows that (MI) is a no-regret method, and that its success rate converges to or outperforms that of the best performing candidate method.

4 Optimal Probability Aggregation

In *probabilistic* prediction games, each candidate method identifies the predicted real value with its credence of the predicted event. We expand the framework from above: Now it contains a series of events represented by random variables $\mathbf{E}_1, \mathbf{E}_2, \dots$ within a space of discrete, mutually disjoint, and exhaustive values $v_1, \dots, v_k$. In order to indicate which value a random variable took on at a specific round, we assume a valuation function *val* to be given by $val_t(v_m) = 1$ if the value of $\mathbf{E}_t$ is v_m and $val_t(v_m) = 0$ otherwise. Predictions are the credences of n [357] candidate methods for each event variable $\mathbf{E}_t$ in the series, represented by probability distributions $\mathbf{P}_1, \dots, \mathbf{P}_n$ such that $\sum_{m=1}^{k} \mathbf{P}_{i,t}(v_m) = 1$ and $\mathbf{P}_{i,t}(v_m) \geq 0$. The probabilistic meta-inductive method $\mathbf{P}_{mi}$ is also represented by a probability distribution. In order to define it, we average the success-rates for the individual values of the value space. Let us first define such an average loss measure l_{av}:

$$l_{i,t}^{av} = \frac{\sum\limits_{m=1}^{k} l(\mathbf{P}_{i,t}(v_m), val_t(v_m))}{k}$$

Note that if l is the quadratic loss function, then l^{av} is the Brier score for a particular round (Brier 1950). The general Brier score can be calculated then by summing up all the scores up to round t and dividing them by t (that is the per round loss averaged over all values of the value space). Now we can define a measure for average success $s_{i,t}^{av}$ in analogy to s (simply replace l by l^{av} in the definition of $s_{i,t}$ above). Likewise, we define weights $w_{i,t}^{av}$ for the probabilistic predictions (simply replace s by s^{av} in the definition of $w_{i,t}$ above). Finally, we define the meta-inductive method for weighted average probability aggregation based on these weights in accordance with (AM): $\mathbf{P}_{mi,t+1} = \sum_{i=1}^{n} w_{i,t}^{av} \cdot \mathbf{P}_{i,t+1}$. Since we assumed that l is convex, also l^{av} is convex. To recognize this, we just have to hint to the mathematical fact that if the loss function l is convex with respect to all values of the value space, then also averaging among the losses with respect to all values of the value space is convex. Since the definition of $\mathbf{P}_{mi}$ is an instance of (MI), and since l^{av} used to determine the weights w^{av} is convex, we can transfer the no-regregt optimality result of P_{mi} to $\mathbf{P}_{mi}$ straightforwardly:

Theorem 2. *Given the loss function l is convex it holds:*

$$s^{av}_{i,t} - s^{av}_{mi,t} \leq \sqrt{n/t} \ \forall i \in \{1, \ldots, n\}, \text{ so } \lim_{t\to\infty} max(s^{av}_{1,t}, \ldots, s^{av}_{n,t}) - s^{av}_{mi,t} \leq 0$$

To conclude: Success-based weighting allows for optimal probability aggregation.

References

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1, pp. 1–3.

Cesa-Bianchi, Nicolo and Lugosi, Gabor (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511546921.

Feldbacher-Escamilla, Christian J. and Schurz, Gerhard (2020b). "Optimal Probability Aggregation Based on Generalized Brier Scoring". In: *Annals of Mathematics and Artificial Intelligence* 88.7, pp. 717–734. DOI: 10.1007/s10472-019-09648-4.

List, Christian and Pettit, Philip (2002). "Aggregating Sets of Judgments: An Impossibility Result". In: *Economics and Philosophy* 18.01, pp. 89–110.

Rossi, Francesca, Venable, Kristen Brent, and Walsh, Toby (2011). *A Short Introduction to Preferences. Between Artificial Intelligence and Social Choice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Williston: Morgan & Claypool.

Schurz, Gerhard (2008). "The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem". In: *Philosophy of Science* 75.3, pp. 278–305. DOI: 10.1086/592550.

— (2019). *Hume's Problem Solved. The Optimality of Meta-Induction*. Cambridge, Massachusetts: The MIT Press.