Project Update: CS281, Fall 2017

Brian Hentschel, Anna Sophie Hilgard, Casey Meehan

November 18, 2017

1 Abstract

Classical peer prediction mechanisms for both ground-truth and no ground-truth cases often assume that users are homogeneous - that is, that their signals are drawn from identical distributions. However, in many no ground-truth environments this model is a poor fit for the true nature of the data. A simple motivating example is that of scoring user reviews (of restaurants, movies, etc.) with peer prediction. Truthful opinions can be expected to vary based on individual characteristics, and we prefer not to penalize users for reporting honestly simply because their opinions are less prevalent than others'.

Recent work by Agarwal et al. [2017] showed that mechanisms exist which are informed-truthful within an error bound when clustering users based on similarity of signal confusion matrices. However, the clustering mechanisms described in this paper are chosen for to define a provable bound rather than optimize for empirical performance. We will instead test a variety of clustering mechanisms, potentially both model-based and model free, and attempt to achieve a grouping which is empirically more likely to encourage truthful feedback than this baseline.

Additionally, many of the datasets used in the empirical work section of this paper actually have a ground truth or gold label, allowing for the use of the Dawid-Skene latent type methods [Dawid and Skene, 1979]. The primary dataset we will focus on, that of the Good Judgment Project Global Forecasting Competition, does not have these labels available, requiring the use of clustering schemes more amenable to missing data. However, it would be an interesting extension to see how error bounds increase when using a proper proxy [Witkowski et al., 2017] as a psuedo-label or whether it is possible to back out individual signal confusion matrices from our estimated correlation

matrices using tensor decomposition methods. With these signal confusion matrices, it may be possible to estimate the preserved signal information for any given individual after removing personal reporting biases, which would potentially allow for a novel information gain-based reward scheme in these settings.

2 Background

In the Good Judgment Project Global Forecasting Competition, the crowd-sourced workers are asked to predict the probability of the occurrence of global political events. It has been shown by Ungar et al. [2012] that this task is surprisingly trainable. That is, with proper feedback and instructions, participants can greatly improve their predictions over time, leading to accurate (by Brier Score) approximations when strategically aggregated over the population. Additionally, work by the Good Judgment Project team has shown that frequent updates are one of the best predictors of prediction accuracy. However, due to the nature of the task, natural feedback is not often available. Tasks frequently ask about the probability of an event within a timeframe of several months. Then, particularly in a negative realization of the event, and often for a positive realization, the answer is not available for some time. Witkowski et al. [2017] have shown that it is in general possible to create a useful proxy for the final realization of the event from an aggregation of current predictions, leading us to believe that we should be able to provide some useful intermediate feedback through peer prediction, leading to increased worker engagement and encouraging worker effort. However, in this environment, differences in reports can come both from asymmetric information and heterogeneous reporting policies. It is critical that we control for heterogeneous reporting to avoid penalizing workers who may have an uncommon but truthful piece of information concerning the outcome. If we assume that the signals themselves are drawn from the same distribution (or distributions based on other potentially observable or inferable features such as expertise), we should be able to achieve truthfulness after controlling for heterogeneity of signal reporting.

3 Definitions and Goals

In each of the following definitions let $[n] = \{1, \ldots, n\}$. The goal of this paper will be to introduce clustering mechanisms that incentivize agents to accurately report their predictions (or classifications). The population of

agents is denoted by P and consists of ℓ agents, numbered 1 through ℓ. For each task, an agent receives a signal in the discrete space $[n]$. This space can be naturally discrete, as in characterizing dogs into dog breeds, or artificially discrete, such as when we partition a continuous space into disjoint ranges. Depending on the task, this signal is observed or it can be generated via proxy methods. Regardless, it is assumed that we have some value for the signal given to each agent. Then for agents p, q we have a delta matrix, which is an $n \times n$ matrix with entry i, j corresponding to agent p receiving signal i and agent q receiving signal j. The entry of the matrix is

$$\Delta_{p,q}(i,j) = D_{p,q}(i,j) - D_p(i)D_q(j)$$

where $D_{p,q}(i,j)$ represents the joint probability of agents p, q receiving signals i, j and $D_p(i), D_q(j)$ the marginal probabilities of agents p, q receiving signals i, j. If we abstract away most of the mechanism design aspects, our goal will largely be to cluster the population of agents into some distinct number of clusters such that

1. Within a cluster, the Delta matrices between agents have very small entries.

2. Between clusters a, b, the Delta matrices between any two agents, with one agent in cluster a and the other in cluster b, should be similar.

4 Baselines and Approaches

We intend to first test our methods on datasets with simpler, more interpretable signals like those used to develop the empirical portion of Agarwal et al. [2017]. The Adult Dataset (`https://github.com/ipeirotis/Get-Another-Label`) and HCB Dataset
(`http://ir.ischool.utexas.edu/square/data.html`) both have relatively high upper bounds on the truthfulness for low numbers of clusters and exhibit some degree of personal judgment in their tasks, identifying the sexual content rating of websites and assessing relevance of web results, respectively. These characteristics suggest that these may be more interesting targets for studies of heterogeneity in users than, for example, labeling dog breeds.

We will then move on to testing methods on the Good Judgment Data, available at
`https://dataverse.harvard.edu/dataverse.xhtml?alias=gjp`. We hope to additionally be able to extrapolate results to a new iteration of this forecasting competition, available at `http://gjopen.com`, from which data is

being scraped. Unique challenges in this problem include the difficulty of quantifying the underlying variable and identifying workable grouping techniques for users with sparse data.

The initial steps of our project have involved preprocessing this data into the delta matrices described in Agarwal et al. [2017], on which we hope to be able to cluster the users. We also must develop the regret calculation framework to evaluate the truthfulness measure for each peer prediction mechanism on which we hope to test, beginning with Correlated Agreement. The baselines against which we hope to improve are those given in the final section of the paper, stating the upper bound on truthfulness incentive bound using clustering based on tensor decomposition methods (which require the existence of per-user confusion matrices).

We will also code up the algorithm used in the paper to initially describe the upper-bounded clustering algorithm and then move on to exploring machine-learning based clustering algorithms, likely beginning with those in the clustering chapter of Murphy to see what is most applicable to our problem and how we might be able to adaptively learn clusters that correspond to our unique goal of motivating truthfulness. In particular, in many non-laboratory signal distributions, the signals may not be evenly distributed, and so it seems likely that it will be the case that a machine learning approach could learn on which segments of the matrix it is most critical to focus on matching with highest precision so as to lower the regret in expectation over all signals. Initial thoughts on the formulation of the problem suggest that it may be useful to formulate each user's data as a concatenation of all of its correlation matrices with other users. This could in some sense be viewed as an image of pixels that represents each user.

If we are able to generate cohesive user-clusters for the Good Judgment Dataset from the users' predictions, a next step would be to see if we could have predicted those clusters/reporting behaviors a priori based on personality testing data gathered from each subject at the outset of the project. If we can come up with even a rough degree to which certain personality traits may correspond to certain reporting styles, we can help to mitigate the cold start problem for new users in such a system (or perhaps even predict which users are likely to be more informative than others).

As a related non-clustering problem, there is additionally interest in using machine learning algorithms to learn the proper aggregation method over individual forecasts to generate the best possible overall forecast of a

given event. If we have the time to explore the problem fully, we may also investigate using neural networks or other adaptive models to isolate high-performing aggregation methods.

5 Evaluation

We intend to evaluate the clustering algorithms on the expected regret a user experiences by telling the truth under a given peer prediction scheme. That is, the expected gain in feedback reward a user could have gained by adopting any strategy other than truthfulness. We will begin by focusing on the Correlated Agreement mechanism [Shnayder et al., 2016] but may also test other mechanisms or develop custom variations on mechanisms as we see fit.

References

Arpit Agarwal, Debmalya Mandal, David C Parkes, and Nisarg Shah. Peer prediction with heterogeneous users. 2017.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196. ACM, 2016.

Lyle H Ungar, Barbara A Mellers, Ville Satopää, Philip Tetlock, and Jon Baron. The good judgment project: A large scale test of different methods of combining expert predictions. In *AAAI Fall Symposium: Machine Aggregation of Human Judgment*, 2012.

Jens Witkowski, Pavel Atanasov, Lyle H Ungar, and Andreas Krause. Proper proxy scoring rules. In *AAAI*, pages 743–749, 2017.