Algebraic Evaluators in their Context: An Annotated Bibliography

Andrés Corrada-Emmanuel, Data Engines

These annotations are meant to help the reader place algebraic evaluators in relation to other approaches that have been taken and continue to be pursued by the AI research community to evaluate and monitor AI agents on unlabeled data. Brief comments are made about the contribution of each source to our understanding of evaluation and monitoring. Our narrow focus, however, is mostly on two technical details that serve to highlight the difference between evaluation on finite samples versus evaluation on infinite samples.

- How is the notion of error independence defined?
- What is the computational cost of making the evaluation?

Error independence is the crucial attribute in evaluation of noisy judges on unlabeled data. Therefore, we can use its technical definition to segregate different approaches. This simple litmus test highlights whether probability distributions are central to a method or not. The definition we use is defined by the sample and is discussed in the `DifferentNotionsOfIndependence.md` in this repository. Most of the papers discussed here use the expectation of a distribution to define error independence between classifiers.

The other feature of importance in evaluation on unlabeled data is its speed. The more frequently we can evaluate ensembles of AI algorithms, the safer we can become. In streaming applications, like an Autonomous Vehicle using auditors, near real-time use would be most useful. In the case of the algebraic evaluator in this submission, that evaluation is instantaneous for all practical purposes. Hilbert famously proved that all ideals are finitely generated. The problem of handling correlated polynomials is finite. It will eventually be solved by future researchers. Data Engines currently uses the fully correlated polynomial system for three binary classifiers. Solving these polynomial systems beyond four or five will be more than sufficient for most applications. Most of the distribution methods discussed here are minimizing a function using the EM algorithm. As such, their computation time is linear in the size of the evaluation test

set. This much faster computation time is a decided advantage for algebraic evaluators over distribution based methods in near real-time AI audits.

The inference side of *Wisdom of the Crowds*

Algebraic evaluators never decide what the correct decisions for the unlabeled data are. They can be used to make such a decision less error prone - the idea behind error correction with the core theorem - but they are not using any such decisions to infer the performance of the crowd.

The decision side of *Wisdom of the Crowds* was considered during the French Revolution by the Marquis de Condorcet. His mathematical treatment of the correctness of decisions made by noisy juries is considered the 1st mathematical treatment of voting systems.

The inference side of *Wisdom of the Crowds* is less well-known. Here we review the relevant papers more narrowly focused on the topic of evaluating AI algorithms. The topic of evaluation itself is vast and touches many disciplines. This bibliography must focus on works most relevant to the problem of noisy AI judges on unlabeled data.

Dawid, P., and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* (1979), 20–28

This paper is the first one to consider using the crowd, not to decide, but to evaluate itself. The task they consider is classification and whether we can take a tally of the votes of the members of an ensemble and compute an estimate of their average accuracy identifying the class labels. Their approach used the EM algorithm to minimize a likelihood function. It is clear from their paper that they are not interested in inferring any distribution or its properties. Probability is used as a tool to minimize the likelihood function.

Corrada-Emmanuel, A., and Schultz, H. Geometric precision errors in low-level computer vision tasks. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)* (Helsinki, Finland, 2008), 168–175

My interest in evaluating noisy judges on unlabeled data started with this paper. It considered the question of how to best fuse multiple Digitial Elevation Maps made from aerial images. Modern digital cameras have made this process much cheaper but the quality of the resulting maps and their possible disagreements on overlapping regions raises questions about how to best fuse them. We developed an approach that used techniques from compressed sensing to estimate the precision error of an ensemble of regressors without ground truth. This approach can be seen as the application of algebraic evaluation to the task of regression. Thus, algebraic evaluators have now been developed for two well-known tasks in AI - regression and classification.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research 11*, 43 (2010), 1297–1322

Raykar et al revived the neglected work of Dawid and Skene. Their main contribution was to introduce the use of Bayesian models to estimate sample accuracies. This work also lead to a US patent by the Siemmens corporation. The results are mostly about error independent judges. They consider the task of regression and classification. Their results in classification are mostly for error independent classifiers. Their definition of independence is defined by the expectation of the Bayesian distributions they introduce.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, 692–700

The rise of crowdsourcing platforms like Amazon's MTurk also gave impetus to the issue of evaluating noisy judges. This paper considers the problem of estimating the human judges in these platforms. Cost considerations, however, introduce a new wrinkle on the task evaluating the human judges. Ideally, we would like to parcel out work disjointly to the judges. But how can we evaluate them then? This paper considers the

evaluation task when we only have partial overlap in the judges decisions. They modify Dawid and Skene's EM estimation to include probabilistic priors that then depend on unknown hyperparameters. This leads to a set of smoothed EM updates. Their approach assumes error independence as defined by these priors. They do not discuss how long the computations took.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, 2195–2203

This paper appeared in the same NeurIPS conference as the previous paper. It was also motivated by the problem of crowdsourcing noisy judges. They want a more fine-grained evaluation, not just an average grade on the test. The judges are still assumed to be error independent given unknown Bayesian distributions for their performance. Since they are interested in making better decisions, their model also wraps up the unknown true labels into the inference process. After justifying their optimization process - the minimax entropy principle - they offer experimental results showing some improvement over Dawid and Skene's EM approach. Since they have to do minimization, they must carry out computations over their update equations. They do not discuss how long these computations took.

Parisi, F., Strino, F., Nadler, B., and Kluger, Y. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences 111*, 4 (2014), 1253–1258

This work started a second approach to evaluating classifiers on unlabeled data. Instead of the Bayesian approach started by Raykar et al, they proposed a spectral approach that looked at the rank of an unknown covariance matrix characterizing the classifiers. The rank of this matrix is shown to go to one in the limit of infinite samples. They assume the classifiers are error independent and their decisions come from the same underlying, but unknown probability distribution. Their EM procedure is improved by using their spectral approach to derive a better initial guess than the typical one of majority voting. One interesting discussion is the robustness of their method to the presence of voting cartels in the ensemble of judges - a step toward dealing with correlated judges. No mention is made of the speed of their estimations.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. 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, 1260–1268

The problem with the EM algorithm is that it is only guaranteed to find a local minimum. This makes approaches that use EM sensitive to their initial starting conditions. They must assume some grade for the classifiers at the start. This paper considered under what conditions such approaches could attain a global optimum solution. The paper returns to the estimation approach of Dawid and Skene and reconsiders the initial estimate in light of recent spectral approaches. The classifiers are assumed error independent under Proposition 1 that uses an asymptotic equality for moments of the decisions when the judges are partitioned into three disjoint sets. Their two stage EM process is then shown to be error bound using an ϵ-δ proof. They do not discuss computation times but their experiments are carried out using only 10 EM iterations and most show that results stabilize after 4 iterations or so.

Jaffe, A., Nadler, B., and Kluger, Y. Estimating the accuracies of multiple classifiers without labeled data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, G. Lebanon and S. V. N. Vishwanathan, Eds., Proceedings of Machine Learning Research, PMLR (San Diego, California, USA, 2015), 407–415

Here the spectral approach to evaluation was developed further. Results for error independent classifiers under an unknown probability distribution are given in the ϵ-δ style of proofs that require infinite samples to show convergence in probability in the limit of infinite samples. They also consider the feedback of their inference results to create an unsupervised ensemble learner. One crucial assumption they must make is that more than half the classifiers are better than 50%. No such assumption is needed for algebraic evaluation.

Platanios, E. A., Dubey, A., and Mitchell, T. Estimating accuracy from unlabeled data: A bayesian approach. In *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of Machine Learning Research* (New York, New York, USA, 2016), 1416–1425

This paper continues the approach of using Bayesian distributions to evaluate classifiers on unlabeled data. It develops an even more complicated probabilistic process with its concomitant introduction of multiple hyperparameters. This is necessary to go beyond the error independence assumption. But otherwise, no theoretical justifications for the model are given. The validity of the model is only considered experimentally - it performed better than other Bayesian approaches.

Steinhardt, J., and Liang, P. S. Unsupervised risk estimation using only conditional independence structure. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc. (2016), 3657–3665

This paper is notable because it considers evaluation as a vehicle to minimize the risk incurred by the errors of the noisy classifiers. All errors are not created equal. A realistic policy for risk minimization would thus have to consider how to best balance their noisy opinions so as to minimize future expected risk. The very first sentences call the problem of evaluation on an unlabeled sample the *risk estimation problem*. It is not clear why this equality between tasks estimating two different statistics is established. They remark that risk estimation would be easy if we had a good evaluation method for average performance since the risk formula essentially convolves costs with error classification rates. Many sample statistics are, like risk, calculated on the basis of a simpler sample statistic. Eventually we will identify the few core sample statistics that we can solve well and preferably, have closed algebraic solutions for them. They will be fast and reliable estimators that can serve as inputs for methods like the ones presented in this paper.

Jaffe, A., Fetaya, E., Nadler, B., Jiang, T., and Kluger, Y. Unsupervised ensemble learning with dependent classifiers. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, A. Gretton and C. C. Robert, Eds., Proceedings of Machine Learning Research, PMLR (Cadiz, Spain, 2016), 351–360

The spectral approach to evaluation is extended to dependent classifiers in this work. The confusion between evaluation and learning is increased by using the a mixed terminology - *unsupervised ensemble learning*. They are trying to learn properties of unknown distributions from a sample test. The incoherence of this approach for evaluation can be seen in their discussion over whether we can assume a single probability distribution for the classifier decisions when different items could be easier or harder to classify. A grade for a test is not a statistic that tells you which questions were hard in the test. Why should evaluation need to take into account item difficulty? This problem only arises when you imagine a stochastic process that generated the judges decisions. No such imaginary process could be assumed for human judges. Algebraic evaluation bypasses all these confusions. Since it does not need them, all discussions about what generated the test results are not needed. The price we pay for that clarity is that we just get a grade for the sample. Algebraic evaluators only estimate a single class of sample statistics. They cannot be used, like methods that estimate probability distributions, to compute other statistics or predict performance on future samples.

Zheng, Y., Li, G., Li, Y., Shan, C., and Cheng, R. Truth inference in crowdsourcing: Is the problem solved? In *Proceedings of the VLDB Endowment*, vol. 10, no. 5 (2017)

No is the answer to the rhetorical question in this paper's title. They perform a series of experimental studies implementing various proposed crowd evaluation methods such as the one proposed by Raykar et al. They tested them across different datasets and found that no method generalized well. All the methods tested were based on probability distribution assumptions.

Algebraic Geometry and Statistics

Pistone, G., Riccomagno, E., and Wynn, H. P. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman and Hall, 2001

This is the first book explaining the connection between algebraic geometry and statistics. The authors were principal actors in developing the connections between the two fields. They show how algebraic geometry can help reframe many classical concerns in statistics such as experimental design and inference of distribution parameters. It has no discussion of evaluation of finite samples on unlabeled data. Dawid and Skene's work is not cited.

Sullivant, S. *Algebraic Statistics*. American Mathematical Society, Providence, 2018

This is a modern introduction that incorporates newer developments in the field. For example, algebraic methods are used to consider the problem of inferring hidden variables in a Markov process assumed to produce sequential data such as DNA. It has no discussion of the algebraic geometry of evaluating noisy judges on unlabeled data.

Cox, D., Little, J., and O'Shea, D. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, 2015

A very popular undergraduate textbook that takes a computational approach to Algebraic Geometry instead of the heavily abstract approach that was started by Grothendieck. All the results derived in this submission can be understood by a reader of this book. The purpose of algebraic evaluators is to return grades quickly, so computational concerns dominate any development of new algorithms.

1 A Sample of Current AI Approaches

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety

This paper is the call for papers in the upcoming NeurIPS 2022 Workshop on ML Safety. The paper does a broad survey of open problems in four different areas - Robustness, Alignment, Monitoring, Systemic Safety. The discussion of Monitoring, the

portion most relevant to this submission, focuses on three sub-areas - Anomaly Detection, Representative Model Outputs, and Hidden Functionality. Two observations about the discussion in this section are relevant to the claim that the AI community is mostly focused in just using probability theory for monitoring. All the papers discussed in this section use probability distributions to carry out monitoring tasks. In addition, the earlier work on evaluation with Bayesian and spectral methods that we have been discussing above is completely absent from the references. The role and importance of evaluation for AI auditing is completely absent from a discussion of open problems in ML safety. No citation of Dawid and Skene's article, either.