Algebraic Ground Truth Inference for Possibly Independent Binary Classifiers via Data Moments

Anonymous Authors[1]

Abstract

The evaluation of the performance of binary classifiers on unlabeled data is considered in a purely algebraic fashion free of asymptotic or distributional assumptions. An exact polynomial system is derived that expresses the frequency of observed classifier voting patterns in terms of the desired, but unknown, sample statistics that require knowledge of the true labels. The system is exactly solvable just using three classifiers - if they are independent in their sample errors. The algebraic nature of the solution allows its practical use when four classifiers are used to detect whether the independence assumption is valid. We also consider the case of generally correlated classifiers. Experiments on the Penn ML binary classification benchmarks demonstrate the utility of this algebraic approach.

1. The evaluation task for binary classifiers

Consider the problem of evaluating the performance of an ensemble of binary classifiers on the instances of a given sample. If we had the correct binary label for each instance, what we call the *ground truth*, the evaluation would be trivial. It neither requires any knowledge of the type of classifiers used or any statistical assumptions about how the ensemble made its decisions. Can such an assumption free evaluation be done when we do not have any knowledge of the correct labels for a sample of decisions by an ensemble of classifiers? This is the problem of *ground truth inference* (GTI) for binary classifiers that we consider here from an algebraic point of view. In general, GTI uses the ensemble of the decisions by learners to decide how correct they are - a problem first treated mathematically by Dawid and Skeene in 1979 (Dawid & Skene, 1979). This is to be contrasted with using the ensemble of decisions to estimate the correct label for each instance in a given sample, which was first treated mathematically by the Marquis de Condorcet during the French Revolution.

Our approach is to equate observable data moments to polynomials of the unknown sample statistics that we call *ground truth statistics*. The data moments are the frequencies with which n binary classifiers voted for the 2^n possible n-label lists observed when we align their decisions by test sample instances. The ground truth statistics are the sample statistics that are defined using the correct labels. Here we will be focusing on the true prevalence of the two labels and the marginal label accuracies of the classifiers. Although we will also consider ground truth statistics that measure how non-independent the classifiers are in the sample being evaluated.

Underlying all of the algebraic algorithms is the assumption that the ground truth exists. So, for example, we disregard cases where the sample instances change labels before being presented to each classifier, so we can align their decisions. The fixed nature of instance variables also means we can write all observed voting pattern frequencies as the sum of two polynomials in the ground truth statistics - the cases where the instances where one label or the other.

1.1. Why only sample statistics?

There are practical and theoretical reasons for focusing on sample statistics absent distributional or asymptotic requirements. A GTI algorithm that relies on few theoretical assumptions is going to have wider applicability than one that depends on particular distributions. The black-box approach we take here also means these algorithms can be applied to any intelligent agent, machine or not. A paucity of assumptions also makes this useful for autonomous systems that may encounter new or variable environments where there is no previous knowledge of how classifiers would act. Safety considerations also mean we want to carry out evaluations on the spot. Asymptotic guarantees are not useful in such cases.

The theoretical advantage of considering just sample statistics is that we will be able to formulate an exact polynomial system that describes any ensemble of classifiers. In general, this polynomial system is under-determined. But we will use it to develop tests for the one assumption we do make throughout most of this paper - that the ensemble of classifiers are independent in their sample errors. The the-

oretical utility of this exact polynomial formulation is that we will use it to develop various unsupervised tests for the independence assumption itself. This is done by connecting the algebraic structure of the independent solution to the "semantics" of our application context - all of our ground truth statistics are integer ratios in a bounded interval.

1.2. Relation to previous work

The use of algebraic geometry to elucidate statistical questions dates back to the later part of the 20th century (Pistone et al., 2001) and goes under the name of *Algebraic Statistics*. Our innovation here is to extend the algebraic approach to inferring sample statistics alone.

Dawid and Skeene's pioneering work on GTI (Dawid & Skene, 1979) relied on the EM algorithm for estimating the prevalence and label accuracies of independent binary classifiers. Since the EM algorithm cannot provide strong guarantees for global optima, this approach has not been developed further. One important branch of GTI algorithms for binary classifiers was started by Raykar *et al.* (Raykar et al., 2010) in 2010 when they introduced Bayesian distributions for evaluating classifiers. The Bayesian approach was further developed by Liu *et al.* (Liu et al., 2012) and Zhou *et al.* (Zhou et al., 2012). A theoretical guarantee of convergence the ground truth values with the Bayesian approach was provided Zhang *et al.* (Zhang et al., 2014) in 2014. However, there is empirical evidence that these Bayesian approaches have limited domain applicability as considered by (Zheng et al., 2017).

A more promising approach for pure evaluation was started by Parisi *et al.* (Parisi et al., 2014). They introduced a spectral approach by constructing a covariance matrix for the classifiers that the off-diagonal elements can be modeled by a rank-one matrix. This result depends on probabilistic proofs that show their method for ranking the classifiers is provably correct in the asymptotic limit. They used their spectral approach to define an optimally weighted ensemble - the Spectral Meta-Learner.Our algebraic approach also allows a way to minimize the errors in a sample set but relies on a straightforward comparison of two polynomial terms. We illustrate this approach when we discuss our experimental results and compare error minimization via algebraic GTI versus majority voting.

Jaffe *et al.* (Jaffe et al., 2015) extended the spectral approach further to solve the case we consider here - estimating the prevalence and label accuracies for independent classifiers. They show that in the asymptotic limit their estimates have errors of $O(1/\sqrt{n}$ where n is the size of the sample. Our algebraic approach provides the exact answer when we have three independent binary classifiers. They also consider the multi-class case. Similarly, our algebraic approach can be trivially generalized to the multi-class case. However, it is not clear how useful the algebraic approach is for the multi-class problem since the resulting polynomial systems we have to solve quickly become computationally intractable.

Jaffe *et al.* (Jaffe et al., 2016) continued developing the spectral approach by extending it to the case of dependent classifiers. They too were concerned with the problem of detecting, in an unsupervised manner, when the independence assumption is violated. We also consider that topic here but in an algebraic fashion. We will discuss classifier dependency briefly in our experimental results but it is not the main focus of this paper. We merely consider when the violations are small enough that we can use the independent solution we develop.

Platatois *et al.* are closest to our algebraic approach. They develop a linear system for the label weighted error of each classifier (Platanios et al., 2016). We discuss further the similarities and differences with our approach in the Supplement.

Ahsen *et al.* (Ahsen et al., 2019) have recently paired the spectral approach with knowledge of the scores provided by the classifying algorithms. They also consider the case of dependent classifiers. Having additional side information like classifier scores should help improve GTI. This limits their approach to machine algorithms. The black box approach we take here means that our algorithms could be applied to any ensemble of learners, possibly humans not just machines. It also means that we retain a small memory footprint since we do not have to keep track of instance scores.

We do not mean to imply by continuing to point out there is no probability in the algebraic approach that the above approaches are inferior. Their approach allows us to generalize their estimates to new, unseen samples and can also incorporate causes for the classifier errors - an important consideration when an autonomous AI agent wants to take corrective action. Rather we mean to emphasize that the algebraic approach shown here can be used in conjunction with them. The use of many redundant systems is a common heuristic for increasing the safety of machines. An example of how GTI can be used to minimize risk functions is given by the work of Steinhardt and Liang (Steinhardt & Liang, 2016). They consider the general multi-class problem and rely on a Bayesian approach.

2. The algebra of sample statistics for a single classifier

Many of the aspects of algebraic GTI can be understood with the simple case of a single classifier. Throughout the paper we will use the abstract labels α and β to denote the two classes that the classifiers are trying to identify. The prevalence of the two labels will be denoted by ϕ_α and

ϕ_β respectively. By the definition of prevalence, these two sample statistics must satisfy $\phi_\alpha + \phi_\beta = 1$. We will use this condition to write all our polynomials in terms of just one of the prevalences, we choose ϕ_α arbitrarily.

Now consider the decisions of a single classifier on the sample. The frequency of the α and β decisions also obey a consistency condition, $f_\alpha + f_\beta = 1$. These observable frequencies can be written exactly in terms of the unknown alpha prevalence and the unknown label accuracies of the classifier, $\phi_{i,\alpha}$ and $\phi_{i,\beta}$, as follows

$$0 = -f_\alpha + \phi_\alpha\phi_{i,\alpha} + (1-\phi_\alpha)(1-\phi_{i,\beta}) \quad (1)$$
$$0 = -f_\beta + \phi_\alpha(1-\phi_{i,\alpha}) + (1-\phi_\alpha)\phi_{i,\beta} \quad (2)$$

These equations can easily be solved by hand. But it is relevant to our later discussion to frame their solution in the language of algebraic geometry.[1]

As the name implies, algebraic geometry, is concerned with the algebra and geometry of polynomial systems. The main algebraic object of interest is the *polynomial ideal*. Since these equations are zero, arbitrary multiplications and sums of them by other polynomials will also be zero. All of these polynomial equations define the polynomial ideal for a particular set of polynomials. Hilbert famously proved that, while there are an infinite number of polynomials in an ideal, they can all be expressed by a non-unique, finite set of polynomials (Cox et al., 2015).

The geometry of the polynomial system is expressed by considering the set of points in the variable space (ϕ_α, $\phi_{i,\alpha}$, and $\phi_{i,\beta}$ in this case) that solve the system. These set of points is called a *variety*. Of central concern to us is if the variety of the polynomial system we construct for the sample statistics of binary classifiers is zero-dimensional. We will abuse terminology in this paper by calling systems that have a zero-dimensional variety *identifiable*. The usual meaning of identifiability is the existence of a single point solution. In this strict sense of the word, no GTI algorithm is identifiable as has been noted by many authors (e.g. (Jaffe et al., 2015)). As we will see, the best we can do for independent classifiers is to have two point solutions.

One powerful way to solve polynomial systems is to consider a particular basis for the polynomial ideal called the *Gröbner basis* (Cox et al., 2015). This can be obtained with Buchberger's algorithm (Cox et al., 2015). Application of the algorithm to the system for a single binary classifier yields,

$$-1 + f_\alpha + f_\alpha \quad (3)$$
$$f_\alpha - \phi_\alpha + \phi_\alpha\phi_{i,\alpha} - \phi_{i,\beta} + \phi_\alpha\phi_{i,\beta} \quad (4)$$

[1]An accessible introduction is available in Cox, Little and O'Shea (Cox et al., 2015).

The polynomial system can only have solutions if the frequencies sum to one,as we know they do by the definition of the frequencies. In addition, we get a single polynomial for the ground truth sample statistics. This makes it clear that the problem of ground truth inference for a single binary classifier is not identifiable. Instead, the variety consists of a two-dimensional surface in the three dimensional space of this polynomial system as shown in Figure 1

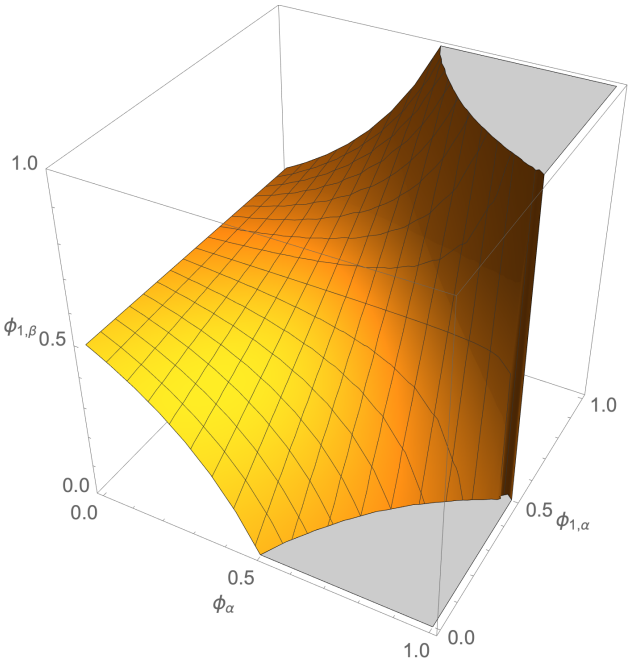

Figure 1. The variety for algebraic ground truth inference of a single binary classifier

This example illustrates that we can derive algebraic expressions for the ground truth statistics using only the observed sample statistics of the frequency of voting decisions. We did not need to assume that the decisions were the result of any particular distribution or process. The rest of the paper considers how far we can take ground truth inference by purely algebraic methods in the variable space consisting of ground truth statistics.

3. Exact polynomial system for arbitrarily correlated binary classifiers

The construction of exact polynomial systems expressing observable ground truth statistics in terms of ground truth sample statistics can be extend to an arbitrary set of classifiers. To do so, we require new ground truth sample statistics that explain how the classifiers are not independent. There are various ways to do this. The one we use here is to define

correlation terms for 2 classifiers for each label, ℓ, by

$$\Gamma_{i,j,\ell} = \frac{1}{n_\ell}\sum_{d=1}^{n_\ell}(\mathbb{1}_{i,d,\ell} - \phi_{i,\ell})(\mathbb{1}_{j,d,\ell} - \phi_{j,\ell}) \quad (5)$$

The indicator functions $\mathbb{1}_{i,d,\ell}$ give 1 if classifier i is correct on instance d for label ℓ and 0 otherwise. This can be generalized to an arbitrary number of classifiers. So we can have m-way label correlation sample statistics where m ranges from 2 to n, the number of classifiers in the ensemble.Independence of classifiers is usually discussed in a probabilistic fashion. Here our definition is just based on these label correlation sample statistics. We call classifiers independent if all these sample correlations are zero.

The correlation sample statistics allow us to write an exact polynomial system for arbitrarily dependent classifiers as stated by the following theorem,

Theorem 1 *Given the set of frequencies $\{f_{\ell_1,\ell_2,\ldots,\ell_n}\}$ of decision events done by an ensemble of n binary classifiers arbitrarily correlated in their sample errors, we can write an exact polynomial system that relates the frequencies to ϕ_α, $\{\phi_{\alpha,i}\}_{i=1}^n$, $\{\phi_{\beta,i}\}_{i=1}^n$ and the error correlations on the sample, $\{\Gamma_{i,j;\ell}\}$ up to $\{\Gamma_{i,j,k,\ldots,n;\ell}\}$.*

We provide the proof in the Supplement. For a general system of n classifiers, the polynomial system contains the following ground truth statistics: one for the α prevalence, $2n$ for the sample accuracies, $n(n-1)$ for the pair correlations, $n(n-1)(n-2)/3$ for the 3-way correlations, and so on up to 2 for the n-way correlations. This sums to up to $2^{n+1}-1$ variables for the ground truth statistics. Since the number of observable frequencies is 2^n of which 2^n-1 are independent, it is clear that there are not enough polynomials for the general problem of GTI to be identifiable with just voting frequencies.

However, if the classifiers are independent in their sample errors, the number of variables remains at $1+2n$. And for $n=3$, we have 7 unknown variables with 7 independent polynomials. As the next section details, this system is identifiable. The utility of Theorem 1 is that it will allow us to develop a purely algebraic test for the possible independence of classifiers on a given sample. We consider the case of independent binary classifiers next.

4. Three independent binary classifiers are identifiable

Our main result for independent binary classifiers is given by the following theorem.

Theorem 2 *The polynomial system for three independent binary classifiers is,*

$$f_{\alpha,\alpha,\alpha} = \phi_\alpha\phi_{1,\alpha}\phi_{2,\alpha}\phi_{3,\alpha} + (1-\phi_\alpha)(1-\phi_{1,\beta})(1-\phi_{2,\beta})(1-\phi_{3,\beta}) \quad (6)$$

$$f_{\alpha,\alpha,\beta} = \phi_\alpha\phi_{1,\alpha}\phi_{2,\alpha}(1-\phi_{3,\alpha}) + (1-\phi_\alpha)(1-\phi_{1,\beta})(1-\phi_{2,\beta})\phi_{3,\beta} \quad (7)$$

$$f_{\alpha,\beta,\alpha} = \phi_\alpha\phi_{1,\alpha}(1-\phi_{2,\alpha})\phi_{3,\alpha} + (1-\phi_\alpha)(1-\phi_{1,\beta})\phi_{2,\beta}(1-\phi_{3,\beta}) \quad (8)$$

$$f_{\beta,\alpha,\alpha} = \phi_\alpha(1-\phi_{1,\alpha})\phi_{2,\alpha}\phi_{3,\alpha} + (1-\phi_\alpha)\phi_{1,\beta}(1-\phi_{2,\beta})(1-\phi_{3,\beta}) \quad (9)$$

$$f_{\beta,\beta,\alpha} = \phi_\alpha(1-\phi_{1,\alpha})(1-\phi_{2,\alpha})\phi_{3,\alpha} + (1-\phi_\alpha)\phi_{1,\beta}\phi_{2,\beta}(1-\phi_{3,\beta}) \quad (10)$$

$$f_{\beta,\alpha,\beta} = \phi_\alpha(1-\phi_{1,\alpha})\,\phi_{2,\alpha}\,(1-\phi_{3,\alpha}) + (1-\phi_\alpha)\phi_{1,\beta}(1-\phi_{2,\beta})\phi_{3,\beta} \quad (11)$$

$$f_{\alpha,\beta,\beta} = \phi_\alpha\phi_{1,\alpha}(1-\phi_{2,\alpha})(1-\phi_{3,\alpha}) + (1-\phi_\alpha)(1-\phi_{1,\beta})\phi_{2,\beta}\phi_{3,\beta} \quad (12)$$

$$f_{\beta,\beta,\beta} = \phi_\alpha(1-\phi_{1,\alpha})(1-\phi_{2,\alpha})(1-\phi_{3,\alpha}) + (1-\phi_\alpha)\phi_{1,\beta}\phi_{2,\beta}\phi_{3,\beta}. \quad (13)$$

It has two point solutions whenever all the voting pattern frequencies, $\{f_{\ell_1,\ell_2,\ell_3}\}$, are greater than zero and not equal. One of the two solutions corresponds to the ground truth value of prevalence of the α label, ϕ_α, and the classifiers accuracies on each of the labels, $\{\phi_{\alpha,i}\}_{i=1}^3$ and $\{\phi_{\beta,i}\}_{i=1}^3$.

The proof provided in the Supplement is a constructive one that relies on Buchberger's algorithm. The main feature of the solution is that the Gröbner basis for the polynomial system contains a single quadratic polynomial for ϕ_α in terms of the observed voting pattern frequencies. The remaining polynomials in the basis are linear equations for each of the unknown accuracies in terms of ϕ_α and the frequencies. Thus there are only two point solutions. The necessity for two solutions is based on the invariance of the polynomial system to the following transformations: $\phi_\alpha \to (1-\phi_\alpha)$, and $\{\phi_{i,\ell} \to (1-\phi_{i,\ell})\}$. The condition that not all frequencies are equal is meant to exclude the case of all three classifiers being 50% accurate on both labels. In that case, all possible values of ϕ_α are solutions to the system. Since equality of the voting frequencies is observable, this condition can be recognized in an unsupervised manner. Jaffe et al. (Jaffe et al., 2015) have also commented on this ambiguity for the multi-class problem.

By themselves, the two point solutions are ambiguous as to which one is the correct one for the ground truth values. In practice, this is not a significant barrier just as error correcting codes are regularly used even though they also have ambiguous solutions. In most practical contexts we can exploit additional side information such as scientific knowledge about the prevalence of the labels or by assuming

that most classifiers are better than 50% . The latter situation is similar to that for error correcting codes where few stray bit flips, not many, are assumed whenever bit flip errors are corrected.

Aside from the independence assumption, which will be the focus of the rest of the paper, this fully solved algebraic solution would be extremely easy to implement in small memory and computational power devices. Like data streaming algorithms, it only requires a small memory data structure, in this case eight integer count registers. And the exact algebraic solution means that the computation can be carried out essentially instantaneously.

5. Unsupervised algebraic tests for the independent errors assumption

We now detail how the algebraic solution of Theorem 2 can be used to test the independence of three classifiers without any knowledge of the correct labels. As noted above, the algebraic solution for the case of three independent classifiers involves a quadratic polynomial in the unknown ϕ_α. But all of our ground truth statistics must be real, bounded integer ratios. This gives us three tests for detecting a violation of the error independence assumption.

The first two tests are obvious and easy to describe. The most severe violation occurs when the quadratic solution for ϕ_α is imaginary. We have observed this case in strongly correlated classifiers. Less severe, but still a violation, is if the prevalence ϕ_α or any of the classifier accuracies lie outside the range $[0, 1]$. We have also observed this experimentally. Both of these violations immediately tell us that the independent assumption cannot possibly describe the ensemble decisions we are evaluating. These two tests cover the properties of being real and bounded for the ground truth statistics.

But what happens when the independent solution returns sensible values for the prevalence and accuracies? In that case we can utilize the following theorem,

Theorem 3 *Given the frequency of voting patterns, $\{f_{\ell_1,\ell_2,\ell_3}\}$, for three binary classifiers, the independent solution quadratic polynomial that solves for the sample prevalence ϕ_α will contain the square root of an irreducible polynomial that is not functionally a perfect square except for the case of zero error correlations in the sample.*

The correct solution for ϕ_α must be an integer ratio. The presence of an irreducible square root is a clear signal the independent model is not strictly correct. We now discuss how Theorem 3 can be proven when we use the exact polynomial formulation given by Theorem 1. A detailed proof is given in the Supplement.

The ϕ_α quadratic is enormous but it can be represented easily in the following symbolic form,

$$a(\{f_{\ell_1,\ell_2,\ell_3}\})\phi_\alpha^2 + b(\{f_{\ell_1,\ell_2,\ell_3}\})\phi_\alpha + c(\{f_{\ell_1,\ell_2,\ell_3}\}) = 0, \quad (14)$$

where the coefficients $a(\{f_{\ell_1,\ell_2,\ell_3}\})$, $b(\{f_{\ell_1,\ell_2,\ell_3}\})$, and $c(\{f_{\ell_1,\ell_2,\ell_3}\})$ are themselves polynomials in the $\{f_{\ell_1,\ell_2,\ell_3}\}$ variables. So the independent polynomial solution for ϕ_α is a simple application of the quadratic formula and will contain a square root of the form $\sqrt{b^2 - 4ac}$. The appearance of a non-reducible square root is a clear signal that the independent model cannot be the correct description of the correlations between the classifiers. But can correlated classifiers return a solution to the independent model quadratic that is an integer ratio? In general, they can not.

Here is where the theoretical advantages of having a full, exact polynomial description for the observed voting frequencies, $\{f_{\ell_1,\ell_2,\ell_3}\}$, as given by Theorem 1, becomes useful. Start by noting that the independent polynomial system can be fed back into the $\phi_\alpha$quadratic itself so we can rewrite $a(\{f_{\ell_1,\ell_2,\ell_3}\})$, $b(\{f_{\ell_1,\ell_2,\ell_3}\})$, and $c(\{f_{\ell_1,\ell_2,\ell_3}\})$as polynomials of the unknown statistics. When we do that the square root term in the quadratic formula solution becomes,

$$\sqrt{b(\{f_{\ell_1,\ell_2,\ell_3}\})^2 - 4a(\{f_{\ell_1,\ell_2,\ell_3}\})c(\{f_{\ell_1,\ell_2,\ell_3}\})} = \sqrt{(1-2\phi_\alpha)^2(\phi_\alpha - 1)^4\phi_\alpha^4(\phi_{1,\alpha} + \phi_{1,\beta} - 1)^4} \sqrt{(\phi_{2,\alpha} + \phi_{2,\beta} - 1)^4(\phi_{3,\alpha} + \phi_{3,\beta} - 1)^4}. \quad (15)$$

Since this is the square root of a perfect square, we obtain a self-consistent result - independent classifiers always have integer ratio solutions for the unknown ϕ_α.

If the classifiers are not independent, then we must put in the full polynomial equations assuming non-zero correlations between the classifiers. Having done so, further algebraic manipulations show that the square root term in the quadratic formula becomes,

$$\left(g_1(\{\phi_{...}\}) + g_2(\{\phi_{...}\}, \{\Gamma_{i,j;\ell}\}, \{\Gamma_{i,j,k;\ell}\})\right)^2 * \left(g_3(\{\phi_{...}\}) + g_4(\{\phi_{...}\}, \{\Gamma_{i,j;\ell}\}, \{\Gamma_{i,j,k;\ell}\})\right). \quad (16)$$

The g polynomials, (g_1, g_2, g_3, g_4), are polynomials in the variables of their arguments. The g_1 and g_4 terms are reducible, and when the error correlations are identically zero, this product term reduces to the one in Equation 15. However, the sum $g_3 + g_4$ is not reducible in general and thus cannot be guaranteed to become a perfect square.

It remains an open question under what conditions some correlated systems can yield a reducible square root factor. Our experimental observations are that even classifiers weakly correlated lead to an irreducible square root term. Theorem

3 only proves that independent systems will always have reducible square root arguments and makes it plausible that most correlated systems also have irreducible square roots.

But must we forgo using the independent solution (Theorem 2) even for weakly correlated classifiers? The next section details how four classifiers can be used to estimate the classifier accuracies by using the independent polynomial system for each of the 4 trios we can create for them.

6. The consistency of independent solutions for four classifiers

It should be intuitively clear that, if a set of four classifiers results in four independent solutions where all ground truth statistics are equal to each other, they must be independent. A proof of this intuition could be obtained by purely algebraic means as we detail in the Supplement. It constructs a 32 polynomial system that we have been unable to solve with Buchberger's algorithm. Although Buchberger's algorithm can be proven to terminate, and thus either prove or disprove our intuition, the algorithm is known to be exponential in space and time (Cox et al., 2015).

What we have observed experimentally and will discuss in our experimental results, is that it is possible to construct 4 classifiers on a sample that gives solutions for each classifier accuracy with less than 2% variation. This makes the algebraic independent solution practical for engineering situations where the classifiers have been designed to be as independent as possible. There are many ways to do this. We can use different algorithms for classification, train on different features, and use different training data.

Another experimental observation we have made is that the independent solutions can have small variations for one label but not the other. In that case, the correlations for the label with smaller correlations are always smaller than for the other label. This raises the possibility of future algebraic algorithms for GTI that proceed in two stages. In the initial stage we use the independent solution and measure the variance in the estimates for the accuracies. If both labels have a variance smaller than a user set value, we are done. But if one label has small variance but the other does not, it may be possible to recover the correlation factors. We discuss this further in the Supplement since we have been unable to resolve one way or the other whether this approach is correct.

7. Experimental results

7.1. Methodology for a single experiment

We tried our best to create nearly independent classifiers for the three datasets shown in this experimental section. The reader is seeing the experiments for which we got closest to the independence condition. Our claim is not that the independent solution is always applicable, but that we can recognize cases in which it is. Our approach for inducing as much independence in the errors as possible was to use three basic techniques: using different classification algorithms, reducing the overlap in training data, and having no or little intersection between classifier feature sets.

We trained four classifiers for an experimental run and then applied the independent errors algebraic solution to each of the 4 possible trios. Thus, each classifier will get three estimates for each of the labels. We then utilized the ground truth values we estimated to minimize the errors in the sample so we could compare our algebraic approach to majority voting and an oracle algorithm we call the *GTI oracle*. This oracle uses the exact ground truth statistics to do error minimization. Thus it allows one to compare how close to zero errors one can get if the full GTI problem was solved for binary classifiers.

Our algebraic approach to minimizing classification errors relies on the property that the polynomials can be expressed as sums of polynomials when one or the other label is correct. Having obtained the ground truth sample statistics, we can calculate how many instances for a given pattern correspond to each correct label. Minimizing errors is achieved by picking the estimated majority correct label for all instances of a given voting pattern. This same procedure can be used for both our estimated GTI values and the GTI oracle values.

The three datasets we use in the experiments come from the Penn ML Benchmarks (Olson et al., 2017). They are `two-norm`, `spambase`, and `mushroom`. The Supplement discusses a fourth dataset - `adult`. The datasets use "0" and "1" for the two binary labels. To maintain consistency with our notation and avoid confusing the labels with the classifier indices, we use α for the "0" label and β for the "1" label.

These experiments lack an application context where we would have side-information that would allow us to pick the right solution out of the two point solutions. We resolve this ambiguity in our experiments by assuming that we know whether ϕ_α is less or greater than 50%.

7.2. `twonorm` experiment

The `twonorm` binary classification benchmark consists of 7,400 items (3703/3697) with 20 features. We divided the features randomly into 4 disjoint sets with 5 features each. Four classifiers were trained, using "off-the-shelf" algorithms provided by the `Mathematica` system: NeuralNetwork, GradientBoostedTrees, NaiveBayes, and LogisticRegression. Each classifier was trained on 1500 items randomly selected from a training set of 2000 (1000/1000).

We then assembled their decisions on the remaining items in the benchmark, the test set, and used these as sole input into the repeated application of Theorem 2.

Our exemplar experiment (Figure 2) shows that the consistency between the recovered values is about 1 percent. As noted, each classifier gets three estimates for each of its label accuracies. On the x-axis we plot the ground truth value and on the y-axis, the recovered value using the independent polynomial system solution. For ease of reference, we include the diagonal line so the reader can see when the recovered value is close to the ground truth one. This experiment is an example where both labels had small error correlations and both recovered estimates were close to the ground truth values for the sample.

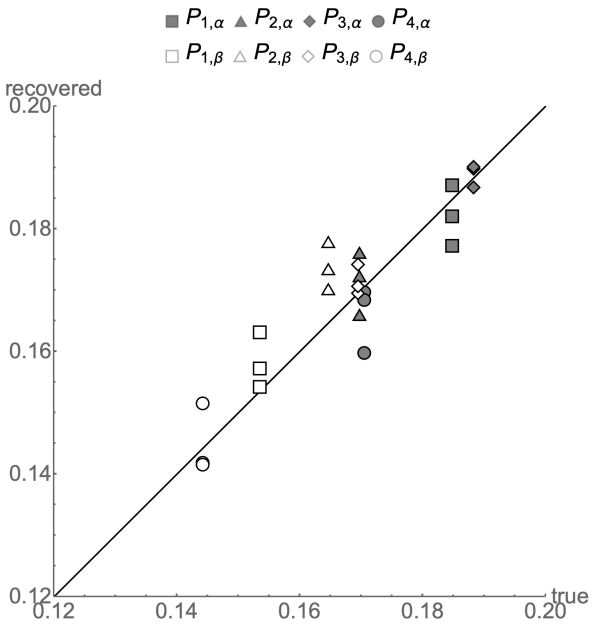

Figure 2. Recovered classifier label accuracy for four classifiers on a single `twonorm` experiment

7.2.1. `twonorm` ERROR MINIMIZATION COMPARISONS

The error minimization results for this experiment (Table 1 highlight the utility of GTI. If we have classifier independence and GTI, accuracy of the classifiers is irrelevant - a meta-learner can just flip the label choices of a classifier that consistently gives the wrong label.

7.3. `spambase` experiment

The `spambase` binary classification benchmark consists of 4601 (2788/1813) items with 57 features. We divided the features randomly into 4 sets of roughly equal size. Four classifiers were trained using the `Mathematica` system algorithms: NeuralNetwork, SupportVectorMachine, DecisionTree and NaiveBayes. Each classifier was trained on (200/100) items randomly selected from a training set of (279/181). These results are an exemplar of how we can recover good estimates for one label with small correlations even in the presence of larger correlations (and higher variance estimates) for the other label.

Table 1. Number of correctly and incorrectly labeled instances in the `two-norm` experiment for various error minimization algorithms

METHOD	CORRECT	INCORRECT
GTI ORACLE	4937	341
GTI (α, β)	4937	341
MAJORITY VOTING	382	4896

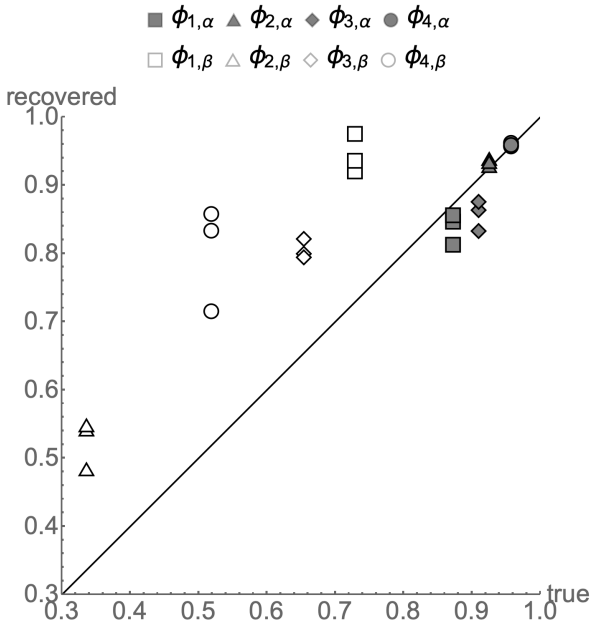

Figure 3. Recovered classifier label accuracy for four classifiers on a single `spambase` experiment

7.3.1. `spambase` ERROR MINIMIZATION COMPARISONS

Although the higher variance of accuracies for the β label alerts us to the inadequacy of these accuracy estimates, we can attempt error minimization because we can estimate the number of β instances for a given voting pattern by subtracting the estimated number of α instances from the total voting pattern instances. We also include the error minimization results when we barrel ahead and use both the α and β independent GTI estimates as shown in Table 2. Surprisingly, this experiment shows that even the weaker GTI estimates in this case are better than majority voting.

Table 2. Number of correctly and incorrectly labeled instances in the `spambase` experiment for various error minimization algorithms

METHOD	CORRECT	INCORRECT
GTI ORACLE	3490	651
GTI (α)	3472	669
GTI (α, β)	3477	664
MAJORITY VOTING	3358	783

7.4. `mushroom` experiment

The `mushroom` binary classification benchmark consists of 8124 (4208/3916) items with 22 features. We divided the features randomly into 4 sets with (6,6,6,5) features each. Four classifiers were trained using the `Mathematica` system algorithms: DecisionTree, NaiveBayes, NeuralNetwork, SupportVectorMachine. Each classifier was trained on (100/100) items randomly selected from a training set of (421/392). These results are an exemplar of roughly equally noisy recovery for both labels.

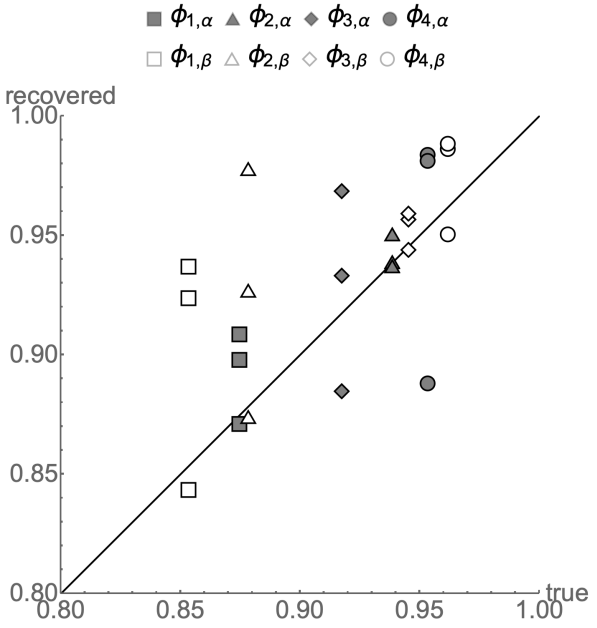

Figure 4. Recovered classifier label accuracy for four classifiers on a single `mushroom` experiment

7.4.1. `mushroom` ERROR MINIMIZATION COMPARISONS

In this experiment both of the labels have high variance estimates. Hence we know that the independent assumption is strongly violated for both labels. Nonetheless, we present comparisons when we pretend the independent GTI estimates are reasonable (Table 3).

Table 3. Number of correctly and incorrectly labeled instances in the `mushroom` experiment for various error minimization algorithms

METHOD	CORRECT	INCORRECT
GTI ORACLE	7088	223
GTI (α, β)	6649	662
MAJORITY VOTING	6909	402

8. Closing remarks

We have presented an algebraic approach for ground truth inference that uses no probability - just the algebra of sample statistics. The approach leads to an exact algebraic solution for the case of independent binary classifiers. We connected the algebraic properties of the independent solution to the "semantics" of our application context - all ground truth statistics are bounded real integer ratios - to develop three unsupervised tests for determining if the independence assumption holds for a set of three classifiers. Since independence in a test sample is unlikely to hold, we considered using four classifiers. Repeated application of the independent classifiers solution allowed us to observe if the classifier errors are close to being independent as assumed. We observed an interesting experimental phenomenon - estimates for one label's accuracies can be close to self-consistent and correct even when the classifiers are more highly correlated in the other label (the `spambase` experiment). In two of our three experimental settings, the GTI estimate allowed us to improve over majority voting when we used GTI to reduce labeling errors in the test sample.

Two features of this algebraic approach make it attractive for enhancing the safety of AI systems. First, the absence of probability assumptions or distributions makes it applicable to more settings where we may not have any idea what is the correct probability distribution that describes classifier errors. Secondly, an algebraic solution, once achieved in general, as we have done here for independent binary classifiers, is a computationally quick estimation method that makes it useful in real-time settings where an AI agent has memory and computational constraints.

Future work is needed to extend these results to more practical settings where we should expect some dependency between classifiers. This algebraic approach is currently computationally hard since it relies on Buchberger's algorithm – a generic algorithm for solving polynomial systems. But the exact polynomial system of Theorem 1 is highly symmetric and other solution methods may be possible. It would also be useful to understand the algebra for the nearly correct recovery of accuracies of an uncorrelated label even as the other label is correlated. Does it hold in general?

References

Ahsen, M. E., Vogel, R. M., and Stolovitzky, G. A. Unsupervised evaluation and weighted aggregation of ranked classification predictions. *Journal of Machine Learning Research*, 20(166):1–40, 2019.

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

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

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

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

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

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(1):36, Dec 2017.

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):1253–1258, 2014.

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

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

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):1297–1322, 2010.

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

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 1260–1268. Curran Associates, Inc., 2014.

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*, volume 10, no. 5, 2017.

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