

---

# Supplement to ‘Streaming algorithms for evaluating noisy judges on unlabeled data - binary classification’

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Introduction

The methodology of using algebraic geometry for finite test evaluation is quite involved and uses mathematics not commonly mentioned in the ML literature. This supplement will try to gently guide the reader to understanding these tools as it provides proofs for the theorem’s mentioned in the paper. It then does a comparison with Theorem 1 from Jaffe et al. for independent classifiers that should be considered the probabilistic counterpart of the algebraic approach used here. Finally, it closes with a detailed discussion of the main experiments in the paper as well as some additional ones.

### 1.1 The general idea of algebraic evaluation

The paper focuses on binary classification but the methodology described is applicable to any number of labels. In addition, it can be used to study decision data sketches for events besides those at the per-item level. This more general framework can be stated as follows -

1. Define the decision event data sketch for the classification stream.
2. Equate each possible decision event to a sum over the true labels.
3. Use the true label indicators to construct exact polynomials describing a label’s contribution to the observed frequency.
4. Compute the set of points in evaluation space, the evaluation variety, that can explain the observed data sketch.

These steps are possible for finite tests because one can always find finite moment expansions for any sample statistic. In essence, we are guaranteed to be able to formulate and prove that a particular polynomial representation can explain all observable data sketches. These polynomial representations may be quite involved but modern computer algebraic systems have no difficulty handling their construction. The last step, finding its corresponding variety is the hard part.

The purpose of an evaluation is to get a ‘grade’ for the ensemble members - an actual number. We will abuse notation by using the same symbols to express the value in an actual test versus the variable used to carry out algebraic formulations. Thus, the prevalence of label  $\alpha$ ,  $\hat{P}_\alpha$ , refers both to its actual value in a test and the variable that defines one of the dimensions in *evaluation space* - the space defined by variables associated to each sample statistic.

### 1.2 The postulate of true or ground truth labels

All the items in the stream have a true label. This ground truth is expressed by the ground truth indicator functions,  $\mathbb{1}_s(\ell)$ .

31 **Definition 1.1.** For each item,  $s$ , in the stream there is a ground truth label indicator function,  $\mathbb{1}_s(\ell)$ ,  
 32 given by,

$$\mathbb{1}_s(\ell) = \begin{cases} 1 & \text{if } \ell = \ell_{\text{true}}^{(s)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

33 The existence of true labels for any one item is then expressed mathematically by,

$$\sum_{\ell_i \in \{\alpha, \beta, \gamma, \dots\}} \mathbb{1}_s(\ell_i) = 1. \quad (2)$$

34 For binary classification there are only two terms in this equation. We could thus choose to represent  
 35 events by polynomials that use variables expressing the frequency of getting a label correct versus  
 36 not. That is, there is only one way to be wrong. But for three or more labels there is more than one  
 37 way to be wrong. In such cases, it would be more natural (i.e. result in more symmetric polynomials)  
 38 if all decision events are expressed in terms of the frequencies of labels wrong. The paper chose to  
 39 describe binary classification events in terms of variables quantifying the frequency of correct label  
 40 decisions.

### 41 1.3 Definitions of the data sketch and its associated sample statistics

42 The black-box approach to noisy judges taken here is that the only information available for evalu-  
 43 ation are observations of their decisions. In the classification task that means that for classifier  $i$  in  
 44 the ensemble we can observe its decisions on the stream items,

$$\{\ell_i^{(s)}\}_{s=1}^n, \quad (3)$$

45 where  $n$  is the number stream items observed so far.

46 **Definition 1.2.** A *per-item ensemble decision event* for item  $s$  is the ordered tuple  $(\ell_1^{(s)}, \ell_2^{(s)}, \dots)$ .

47 From these events we can construct a corresponding data sketch that forgets any information about  
 48 decisions across items and just tallies the per-item decision events.

49 **Definition 1.3.** The *per-item decision data sketch* for an ensemble of  $C$  classifiers is given by the  
 50 integer counters for all possible per-item decision events. For  $L$  possible labels,  $L^C$ , counters are  
 51 needed.

#### 52 1.3.1 The label prevalences

53 The prevalences of the labels in the stream are integer ratios defined using the true label indicator  
 54 functions.

55 **Definition 1.4.** The true prevalence of a label  $\ell$ ,  $\hat{P}_\ell$ , in the observed stream items is given by,

$$\hat{P}_\ell = \frac{1}{n} \sum_{s=1}^n \mathbb{1}_s(\ell) \quad (4)$$

56 It follows from the postulate of ground truth labels that,

$$\sum_{\ell \in \{\alpha, \beta, \gamma, \dots\}} \hat{P}_\ell = 1 \quad (5)$$

57 This equation is not mentioned again here because for binary classification we can make it disappear  
 58 by focusing on just one of the two label prevalences. In general, this would not be possible and this  
 59 equation would be part of the polynomial set that defines the evaluation ideal.

60 Note that when one wants to evaluate performance across stream items, there would be prevalences  
 61 for each of the true label tuples possible. For example, to evaluate accuracies on two consecutive  
 62 stream items, four prevalences would be required during binary classification.

### 63 1.3.2 Classifier label accuracies

64 The performance of an ensemble classifier for  $L$  labels requires as many sample statistics. But by  
 65 Equation 2, one of these could be expressed in terms of the others. In binary classification it is  
 66 sufficient to use the accuracies on each label.

67 **Definition 1.5.** The  $\alpha$  and  $\beta$  accuracies for classifier  $i$  are given by,

$$\hat{P}_{i,\alpha} = \frac{1}{n_\alpha} \sum_{\mathbb{1}_s(\alpha)=1} \mathbb{1}_s(\ell_i^{(s)}) \quad (6)$$

$$\hat{P}_{i,\beta} = \frac{1}{n_\beta} \sum_{\mathbb{1}_s(\beta)=1} \mathbb{1}_s(\ell_i^{(s)}) \quad (7)$$

### 68 1.3.3 Classifier decision correlations

69 The decision correlations during a finite test are defined as products of terms in the form,

$$\mathbb{1}_s(\ell_i^{(s)}) - \hat{P}_{i,\ell}. \quad (8)$$

70 For ensembles of three classifiers we only need two and three way correlation variables for each  
 71 label. These are defined as follows.

72 **Definition 1.6.** The 2-way or pair decision correlation for classifiers  $i$  and  $j$  on label  $\ell$  is given by,

$$\Gamma_{i,j,\ell} = \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} (\mathbb{1}_s(\ell_i^{(s)}) - \hat{P}_{i,\ell})(\mathbb{1}_s(\ell_j^{(s)}) - \hat{P}_{j,\ell}). \quad (9)$$

73 The 3-way decision correlation for classifiers  $i$ ,  $j$ , and  $k$  on label  $\ell$  is expressed as,

$$\Gamma_{i,j,k,\ell} = \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} (\mathbb{1}_s(\ell_i^{(s)}) - \hat{P}_{i,\ell})(\mathbb{1}_s(\ell_j^{(s)}) - \hat{P}_{j,\ell})(\mathbb{1}_s(\ell_k^{(s)}) - \hat{P}_{k,\ell}). \quad (10)$$

### 74 1.3.4 Finite moment expansions

75 Other than carefully defining how these variables are defined in terms of the ground truth indicator  
 76 functions, there is nothing novel or unusual here. Expansions of sample statistics in terms of data  
 77 moments is well-known. For three classifiers we require the first moment variables (the label ac-  
 78 curacies), the second moment variables (the pair correlations) and the third moment variables (the  
 79 3-way correlations).

## 80 2 Theorem 1: the polynomial generating set for independent classifiers

81 It is possible to define mathematical objects that do not exist. Hence, formal mathematical treat-  
 82 ments start by establishing that they exist - existence theorems. Another class of theorems are about  
 83 completeness of representations - that we have a way of describing all possible objects. For exam-  
 84 ple, the completeness of Fourier series in describing all possible piecewise continuous functions.  
 85 Theorem 1 is a combination of these two types of theorems. It proves that the decision data sketch  
 86 of independent classifiers is exactly described by polynomials using only the variables in the *basic*  
 87 *set* of statistics:  $\hat{P}_\alpha$ ,  $\{\hat{P}_{i,\alpha}\}_{i=1}^3$ , and  $\{\hat{P}_{i,\beta}\}_{i=1}^3$ . But because the proof is constructive and starts from  
 88 the true evaluation point, it also proves that the constructed polynomials do define a variety that is  
 89 non-trivial (not the empty set).

90 **Theorem 1.** Each of the decision event frequencies,  $f_{\ell_1, \ell_2, \ell_3}$ , derived from the per-item data sketch  
 91 of three independent classifiers is given by a polynomial in the basic set of statistics. They are the

92 *generating set* of the evaluation ideal for independent classifiers,

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

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

$$f_{\alpha,\beta,\alpha} = \hat{P}_\alpha \hat{P}_{1,\alpha} (1 - \hat{P}_{2,\alpha}) \hat{P}_{3,\alpha} + (1 - \hat{P}_\alpha)(1 - \hat{P}_{1,\beta}) \hat{P}_{2,\beta} (1 - \hat{P}_{3,\beta}) \quad (13)$$

$$f_{\beta,\alpha,\alpha} = \hat{P}_\alpha (1 - \hat{P}_{1,\alpha}) \hat{P}_{2,\alpha} \hat{P}_{3,\alpha} + (1 - \hat{P}_\alpha) \hat{P}_{1,\beta} (1 - \hat{P}_{2,\beta}) (1 - \hat{P}_{3,\beta}) \quad (14)$$

$$f_{\beta,\beta,\alpha} = \hat{P}_\alpha (1 - \hat{P}_{1,\alpha}) (1 - \hat{P}_{2,\alpha}) \hat{P}_{3,\alpha} + (1 - \hat{P}_\alpha) \hat{P}_{1,\beta} \hat{P}_{2,\beta} (1 - \hat{P}_{3,\beta}) \quad (15)$$

$$f_{\beta,\alpha,\beta} = \hat{P}_\alpha (1 - \hat{P}_{1,\alpha}) \hat{P}_{2,\alpha} (1 - \hat{P}_{3,\alpha}) + (1 - \hat{P}_\alpha) \hat{P}_{1,\beta} (1 - \hat{P}_{2,\beta}) \hat{P}_{3,\beta} \quad (16)$$

$$f_{\alpha,\beta,\beta} = \hat{P}_\alpha \hat{P}_{1,\alpha} (1 - \hat{P}_{2,\alpha}) (1 - \hat{P}_{3,\alpha}) + (1 - \hat{P}_\alpha)(1 - \hat{P}_{1,\beta}) \hat{P}_{2,\beta} \hat{P}_{3,\beta} \quad (17)$$

$$f_{\beta,\beta,\beta} = \hat{P}_\alpha (1 - \hat{P}_{1,\alpha}) (1 - \hat{P}_{2,\alpha}) (1 - \hat{P}_{3,\alpha}) + (1 - \hat{P}_\alpha) \hat{P}_{1,\beta} \hat{P}_{2,\beta} \hat{P}_{3,\beta} \quad (18)$$

93 This generating set defines a non-empty evaluation variety that contains the true evaluation point.

94 *Proof.* By the existence of true labels, it follows that any decision event by the classifiers has a count  
95 equal to the sum of times the true labels were  $\alpha$  plus those were the true labels were  $\beta$ ,

$$n_{\ell_1, \ell_2, \ell_3} = \#(\ell_1, \ell_2, \ell_3 | \alpha) + \#(\ell_1, \ell_2, \ell_3 | \beta). \quad (19)$$

96 Dividing this equation by  $n$ , the number of items classified and then multiplying each label term by  
97 unity in the form  $n_\ell/n_\ell$ , this becomes

$$f_{\ell_1, \ell_2, \ell_3} = \frac{n_\alpha}{n} \left( \frac{1}{n_\alpha} \#(\ell_1, \ell_2, \ell_3 | \alpha) \right) + \frac{n_\beta}{n} \left( \frac{1}{n_\beta} \#(\ell_1, \ell_2, \ell_3 | \beta) \right) \quad (20)$$

$$= \hat{P}_\alpha \left( \frac{1}{n_\alpha} \#(\ell_1, \ell_2, \ell_3 | \alpha) \right) + \hat{P}_\beta \left( \frac{1}{n_\beta} \#(\ell_1, \ell_2, \ell_3 | \beta) \right) \quad (21)$$

$$= \hat{P}_\alpha \left( \frac{1}{n_\alpha} \#(\ell_1, \ell_2, \ell_3 | \alpha) \right) + (1 - \hat{P}_\alpha) \left( \frac{1}{n_\beta} \#(\ell_1, \ell_2, \ell_3 | \beta) \right) \quad (22)$$

98 The construction of the generating set then proceeds by reformulating the number of times a decision  
99 event occurred given the true label in terms of the label accuracies of the classifiers. This is tedious  
100 but straightforward.

101 Consider the decision event  $(\alpha, \beta, \alpha)$ . When the true label is  $\alpha$  the number of times this event  
102 occurred is equal to,

$$\#(\alpha, \beta, \alpha | \alpha) = \sum_{\mathbb{1}_s(\alpha)=1} \mathbb{1}_s(\ell_1^{(s)}) (1 - \mathbb{1}_s(\ell_2^{(s)})) \mathbb{1}_s(\ell_3^{(s)}). \quad (23)$$

103 Correspondingly, the composite indicator function for  $(\alpha, \beta, \alpha)$  events when the true label is  $\beta$  is  
104 given by,

$$\#(\alpha, \beta, \alpha | \beta) = \sum_{\mathbb{1}_s(\alpha)=1} (1 - \mathbb{1}_s(\ell_1^{(s)})) \mathbb{1}_s(\ell_2^{(s)}) (1 - \mathbb{1}_s(\ell_3^{(s)})). \quad (24)$$

105 The proof now hinges on whether we can write averages of products of the true indicator functions  
106 as products of their averages when the classifiers are independent.

107 Every decision event composite indicator function like those in Equations 23 and 24 will contain at  
108 most pair products,

$$\mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}), \quad (25)$$

109 and triple products,

$$\mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}) \mathbb{1}_s(\ell_k^{(s)}). \quad (26)$$

110 Let us look at the pair product using the definition for the pair correlation variables and go through  
111 simplifications that should be familiar,

$$\Gamma_{i,j,\ell} = \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} (\mathbb{1}_s(\ell_i^{(s)}) - \hat{P}_{i,\ell}) (\mathbb{1}_s(\ell_j^{(s)}) - \hat{P}_{j,\ell}) \quad (27)$$

$$= \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}) \right) - \hat{P}_{i,\ell} \hat{P}_{j,\ell}. \quad (28)$$

112 Setting the pair correlation variables to zero then guarantees that averages of products are products  
 113 of averages. In addition, it forms part of the definition of what ‘independence’ means when talking  
 114 about sample statistics. Independent classifiers must have all pair correlation values,  $\Gamma_{i,j,\ell}$ , set to  
 115 zero.

116 The triple product is more involved but follows accordingly from the definition of the 3-way corre-  
 117 lation values and the pair ones,

$$\begin{aligned} \Gamma_{i,j,k,\ell} &= \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} (\mathbb{1}_s(\ell_i^{(s)}) - \hat{P}_{i,\ell})(\mathbb{1}_s(\ell_j^{(s)}) - \hat{P}_{j,\ell})(\mathbb{1}_s(\ell_k^{(s)}) - \hat{P}_{k,\ell}) \\ &= \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}) \mathbb{1}_s(\ell_k^{(s)}) \right) - \hat{P}_{i,\ell} \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_j^{(s)}) \mathbb{1}_s(\ell_k^{(s)}) \right) - \\ &\quad \hat{P}_{j,\ell} \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_k^{(s)}) \right) - \hat{P}_{k,\ell} \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}) \right) + \\ &\quad 2\hat{P}_{i,\ell}\hat{P}_{j,\ell}\hat{P}_{k,\ell} \end{aligned} \quad (29)$$

$$(30)$$

118 Repeatedly invoking Equation 28, the final expression for the 3-way correlation values is,

$$\begin{aligned} \Gamma_{i,j,k,\ell} &= \left( \frac{1}{n_\ell} \sum_{\mathbb{1}_s(\ell)=1} \mathbb{1}_s(\ell_i^{(s)}) \mathbb{1}_s(\ell_j^{(s)}) \mathbb{1}_s(\ell_k^{(s)}) \right) + \hat{P}_{i,\ell}\Gamma_{j,k,\ell} + \hat{P}_{j,\ell}\Gamma_{i,k,\ell} \\ &\quad + \hat{P}_{k,\ell}\Gamma_{i,j,\ell} - \hat{P}_{i,\ell}\hat{P}_{j,\ell}\hat{P}_{k,\ell} \end{aligned} \quad (31)$$

119 This defines the final condition for the sample independence of three classifiers - both  $\Gamma_{i,j,k,\alpha}$  and  
 120  $\Gamma_{i,j,k,\beta}$  are equal to zero. This completes the proof that the generating polynomial set in the theorem  
 121 is sufficient to explain all decision data sketches by independent classifiers using only polynomials  
 122 of the basic set of evaluation variables.

123 The proof quickly concludes by observing that the constructive part of the proof means that we have  
 124 at least one point in the evaluation space defined by the basic set that satisfies all these equations  
 125 when we go from thinking about them as actual values to treating them as variables. The set of  
 126 points that satisfy the equations for a polynomial set is called the *variety* or as is being called here -  
 127 the *evaluation variety*. It is non-empty and it contains the true evaluation point.  $\square$

### 128 3 Theorem 2: the evaluation variety for independent classifiers

129 Theorem 1 constructed the polynomial generating set for independent classifiers and showed that  
 130 the true evaluation point is contained in its variety. Theorem 2 now considers the question of how  
 131 many points, besides the true evaluation point, could also explain an observed decision data sketch  
 132 from independent classifiers.

133 **Theorem 2.** The evaluation variety of the independent classifiers generating set contains exactly  
 134 two points, one of which is the true evaluation point.

135 *Proof.* Solving the generating set for independent classifiers is quite involved algebraically. The  
 136 accompanying Mathematica notebook foo details the calculations. Here we describe the general  
 137 strategy of the proof and add more explanations for the terms involved in the independent evaluator’s  
 138 expression for  $\hat{P}_\alpha$  that appears in Table 1 in the paper.

139 A strategy for solving multi-variable polynomial system is to obtain an equivalent representation of  
 140 the polynomials that create an elimination ‘ladder’. At the bottom of the ladder is a single variable  
 141 polynomial that, by the fundamental theorem of algebra, has as many roots (counting multiplicity) as  
 142 the order of the polynomial. One then climbs up the ladder by finding polynomials that involve the  
 143 solved variable and one more variable. In this manner, one systematically solves for the unknown  
 144 values for the variables satisfying the polynomial system. This alternative representation is called  
 145 the *elimination ideal*. It can be obtained by solving for the Gröebner basis of the generating set. The

$4 \Delta_{1,2} \Delta_{1,3} \Delta_{2,3} + (f_{1,\beta} f_{2,\beta} f_{3,\beta} + f_{3,\beta} \Delta_{1,2} + f_{2,\beta} \Delta_{1,3} + f_{1,\beta} \Delta_{2,3} - f_{\beta,\beta,\beta})^2$
$4 f_{\alpha,\beta,\beta} f_{\beta,\alpha,\beta} f_{\beta,\beta,\alpha} + f_{3,\beta}^2 f_{\beta,\beta,\alpha}^2 + 4 f_{\alpha,\beta,\beta} f_{\beta,\alpha,\beta} f_{\beta,\beta,\alpha} - 2 f_{3,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\beta} + 2 f_{3,\beta}^2 f_{\beta,\beta,\alpha} f_{\beta,\beta,\beta} + 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\beta} +$ $4 f_{\beta,\alpha,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\beta} + f_{\beta,\beta,\beta}^2 - 2 f_{3,\beta} f_{\beta,\beta,\beta}^2 + f_{3,\beta}^2 f_{\beta,\beta,\beta}^2 + 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\beta}^2 + 4 f_{\beta,\alpha,\beta} f_{\beta,\beta,\beta}^2 + 4 f_{\beta,\beta,\alpha} f_{\beta,\beta,\beta}^2 +$ $4 f_{\beta,\beta,\beta}^3 + f_{1,\beta}^2 (f_{\alpha,\beta,\beta} + f_{\beta,\beta,\beta})^2 + f_{2,\beta}^2 (f_{\alpha,\beta,\beta} + f_{\beta,\beta,\beta})^2 - 2 f_{2,\beta} (f_{\alpha,\beta,\beta} + f_{\beta,\beta,\beta}) (f_{\beta,\beta,\beta} + f_{3,\beta} (f_{\beta,\beta,\beta} + f_{\beta,\beta,\beta})) -$ $2 f_{1,\beta} (f_{2,\beta} (f_{\alpha,\beta,\beta} (f_{\beta,\beta,\beta} + f_{\beta,\beta,\beta}) + f_{\beta,\beta,\beta} (-2 f_{3,\beta} + f_{\beta,\alpha,\beta} + f_{\beta,\beta,\beta})) + (f_{\alpha,\beta,\beta} + f_{\beta,\beta,\beta}) (f_{\beta,\beta,\beta} + f_{3,\beta} (f_{\beta,\beta,\beta} + f_{\beta,\beta,\beta})))$
$f_{\alpha,\beta,\beta}^2 f_{\beta,\beta,\alpha}^2 - 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + f_{\alpha,\beta,\beta}^2 f_{\beta,\beta,\alpha}^2 - 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} +$ $4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta}^2 f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} -$ $4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta}^2 f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} - 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta}^2 f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} - 4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} - 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha}^2 -$ $2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 4 f_{\alpha,\beta,\beta} f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha}^2 - 2 f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} +$ $2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} +$ $4 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha} f_{\beta,\beta,\alpha}^2 f_{\beta,\beta,\alpha} -$ $2 f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 + 2 f_{\alpha,\beta,\beta} f_{\beta,\beta,\alpha}^3 -$

Figure 1: Stages, from the bottom up, of the algebraic simplification of the ‘a’ coefficient in the prevalence quadratic (Eq. 32). The bottom cell shows the expression using the eight data sketch frequencies. The middle cell shows the simplification after introducing the  $\beta$  label decision frequencies for each classifier,  $f_{i,\beta}$ . The top cell was obtained by defining new variables,  $\Delta_{i,j}$ , as explained in the main text.

term ‘basis’ is somewhat unfortunate since it may mislead the reader into thinking that a polynomial basis is like a vector basis. It is not. For example, the number of polynomials in a given basis for an evaluation ideal need not be the same as that in another basis. There is no concept of dimension when talking about evaluation ideals.

Buchberger’s algorithm is a generic algorithm for finding the many Gröbner bases one may care to obtain. The Gröbner basis used in the notebook computations for this proof is also an elimination ideal. It contains at the bottom a quadratic polynomial in  $\hat{P}_\alpha$  of the form,

$$a(\dots)\hat{P}_\alpha^2 + b(\dots)\hat{P}_\alpha + c(\dots) = 0. \quad (32)$$

By the fundamental theorem of algebra, this has two roots. And by Theorem 1 we know that one must be the true prevalence for the  $\alpha$  items. Because the elimination ideal also contains linear equations relating a single  $\hat{P}_{i,\ell}$  variable with  $\hat{P}_\alpha$ , it also follows that each root of the  $\hat{P}_\alpha$  polynomial is associated with a single value for all the  $\hat{P}_{i,\ell}$  variables. This proves that the evaluation variety for independent classifiers contains exactly two points in evaluation space.

The two points are related. Given a point that solves the independent generating set, the other one is given by the transformations

$$\hat{P}_\alpha \rightarrow (1 - \hat{P}_\alpha) \quad (33)$$

$$\hat{P}_{i,\alpha} \rightarrow (1 - \hat{P}_{i,\beta}) \quad (34)$$

$$\hat{P}_{i,\beta} \rightarrow (1 - \hat{P}_{i,\alpha}) \quad (35)$$

160

□

The ambiguity of the exact solution for independent classifiers has been noted before in the literature. It is inherent to all inverse problems and can be found in many fields such as tomography and error-correcting codes.

The raw expression obtained by solving Equation 32 is unwieldy and would be impossible to present here. It would also hide some of the structure that is relevant to understanding the limitations of the formula and of the algebraic approach in general. In the accompanying Mathematica code (?), the necessary operations are carried out. For example, Figure 3 shows three stages of the simplifications of the ‘a’ coefficient in the prevalence quadratic (Equation 32). The top cell was obtained by

introducing new variables. The  $f_{i,\beta}$  variables are the decision frequencies by the classifier  $i$  for the  $\beta$  label. These actually define ‘blind spots’ in the independent evaluator as will be discussed during the proof of Theorem 3. The second set of new variables, the  $\Delta_{i,j}$ , are defined as follows,

$$\Delta_{i,j} = f_{i,j,\beta} - f_{i,\beta} f_{j,\beta}. \quad (36)$$

The new decisions frequencies,  $f_{i,j,\beta}$ , keep track of how often the pair (i,j) both voted for label  $\beta$  on an item. Interestingly, it can be proven that

$$f_{i,j,\beta} - f_{i,\beta} f_{j,\beta} = f_{i,j,\alpha} - f_{i,\alpha} f_{j,\alpha}. \quad (37)$$

So there is no need to define  $\Delta$  variables for each label.

Finally, it should be noted that the 3 independent classifier solution solves the evaluation problem for any ensemble of three or more of them. This follows from marginalizing classifiers out. In addition, it can also be shown that 3 is the minimum number needed for this. The two and one independent classifiers have evaluation variety surfaces, not points.

#### 4 Theorem 3: the generating set for correlated classifiers

The utility of working with sample statistics comes to the forefront in Theorem 3. Algebraic evaluation is easy in comparison to predicting. As will be shown by the theorem, we can write a generating set for data sketches of correlated classifiers that is complete. This is a universal representation of all data sketches no matter the setting. There is no theory of the phenomena that the classifiers analyzed in algebraic evaluation. How can there be when it has no free parameters? Because we are estimating sample statistics, we can write complete representations. There are no unknown unknowns in evaluation because of this.

**Theorem 3.** There is a polynomial generating set for all per-item data sketches by correlated classifiers. It requires the basic set of evaluation statistics  $\hat{P}_\alpha$ ,  $\{\hat{P}_{i,\alpha}\}_{i=1}^n$ ,  $\{\hat{P}_{i,\beta}\}_{i=1}^n$ . In addition, it uses  $n$  choose two pair correlation variables per label,  $n$  choose three 3-way correlation variables and so on until it terminates at 1  $n$ -way correlation variable. It defines an evaluation variety that contains at least one point - the true evaluation point.

*Sketch of the proof.* The proof is constructive as in Theorem 1. We are starting at the true evaluation point. This point now exists in a space of dimension,

$$1 + 2n + 2 \sum_{m=2}^n \binom{n}{m} = 2^{n+1} - 1. \quad (38)$$

The frequency of times any decision event occurs given the true label can be written as the average of a composite indicator function as was done in the proof of Theorem 1. For  $n$  classifiers, the largest product term possible in this composite indicator function would contain  $n$  indicator functions. But these can be reformulated in terms of the basic evaluation statistics, the  $n$ -way correlation variable for the label and averages of products with  $n - 1$  indicator functions. Continuing in this way, the products of the  $n - 1$  can be rewritten in terms of the  $(n-1)$ -way correlation variables and products with  $n - 2$ . This descent ends at  $n = 1$ .

It follows trivially by the constructive nature of the proof that this generating set defines an evaluation variety that contains at least one point - the true evaluation point.  $\square$

The Mathematica notebook `GeneratingPolynomialsCorrelatedClassifiers.nb` contains explicit formulations of the polynomial generating set for two and three classifier ensembles. One practical utility of these general polynomial sets is that they allow theoretical study of the properties of functions based on the data sketch frequencies. In particular, one can use these polynomial expressions when correlation exists to study the behavior of the independent evaluator estimates. Consider the algebraic expression for  $\hat{P}_\alpha$  under the independence assumption. It is an algebraic function of the data sketch frequencies,

$$\frac{1}{2} - \frac{1}{2} \frac{(f_{\beta,\beta,\beta} - (f_{1,\beta} f_{2,\beta} f_{3,\beta} + f_{1,\beta} \Delta_{2,3} + f_{2,\beta} \Delta_{1,3} + f_{3,\beta} \Delta_{1,2}))}{\sqrt{4 \Delta_{1,2} \Delta_{1,3} \Delta_{2,3} + (f_{\beta,\beta,\beta} - (f_{1,\beta} f_{2,\beta} f_{3,\beta} + f_{1,\beta} \Delta_{2,3} + f_{2,\beta} \Delta_{1,3} + f_{3,\beta} \Delta_{1,2}))^2}}. \quad (39)$$

Self-consistency under the independence assumption means that substituting the independent generating polynomials for the frequencies into this expression will result in two solutions,

$$\hat{P}_\alpha, 1 - \hat{P}_\alpha. \quad (40)$$

But the independent estimate is not correct for correlated classifiers. By using the generating set for such classifiers, a Taylor expansion on the variables  $\Gamma_{i,j,\ell}$  can be done for the  $\hat{P}_\alpha$  estimate. This work is not presented here except to remark that the first order terms have the generic form,

$$\frac{\Gamma_{i,j,\ell}}{\hat{P}_{k,\alpha} + \hat{P}_{k,\beta} - 1}. \quad (41)$$

Therefore, the independent evaluator estimates errors grow as one approaches the line,

$$\hat{P}_{k,\alpha} + \hat{P}_{k,\beta} - 1, \quad (42)$$

in evaluation space. The origin of this ‘blind spot’ and others in algebraic evaluation will become clearer in the next section where a Gröebner basis for correlated classifiers is discussed.

## 5 Theorem 4: A Gröebner basis for three correlated classifiers

The generating set for arbitrarily correlated classifiers quickly overwhelms current computational commutative algebra platforms such as Mathematica. They use Buchberger’s algorithm, a generic algorithm that is proven to always terminate in finite time but is known to take exponential time and memory for some polynomial systems, Cox et al. [2015]. Nonetheless, we can get a glimpse of how a general solution looks by solving for Gröebner bases for ensembles of three correlated classifiers. It is this basis that allows us to define a containing variety, larger than the evaluation variety, that contains the true evaluation point.

More importantly, the polynomials in the basis are simplified by expressing them in terms of new variables,  $\pi_{i,\ell}$  and  $\gamma_{i,j,\ell}$ , that are shifted versions of  $\hat{P}_{i,\ell}$  and  $\Gamma_{i,j,\ell}$  respectively,

$$\pi_{i,\ell} = \hat{P}_{i,\ell} - f_{i,\ell} \quad (43)$$

$$\gamma_{i,j,\ell} = \Gamma_{i,j,\ell} - \Delta_{i,j}. \quad (44)$$

**Theorem 4.** A Gröebner basis for the generating set of three correlated classifiers exists. A subset of the basis consists of polynomials of the forms,

$$\hat{P}_\alpha \pi_{i,\alpha} - \hat{P}_\beta \pi_{i,\beta} \quad (45)$$

$$\pi_{i,\alpha} \pi_{j,\beta} - \pi_{i,\beta} \pi_{j,\alpha}. \quad (46)$$

These define a *containing variety* guaranteed to contain the true evaluation point.

*Proof.* The proof is constructive and is given in GroebnerBasis3CorrelatedClassifiers.nb. The inclusion of the true evaluation point in the containing variety follows from the general theorem that a subset of generating polynomials must define a variety that includes the variety of the full set. That is, the set of points that satisfy a subset of the polynomials has to be equal to or larger than the set of points that satisfy all the polynomials.  $\square$

It is not proven here that the containing variety has dimension  $n + 1$  in the  $2n + 1$  space of the basic evaluation statistics of  $n$  correlated classifiers. Since the containing variety does not require any knowledge of the correlation variables, this dimensionality reduction is universally available and can serve as a constraint for other methods such as the probabilistic approaches to evaluation.

## 6 Theorem 5: Unresolved square roots signal correlation but seemingly correct estimates are no guarantee of sample independence

Theorem 5 combines theorems 2 and 3 to prove that unresolved square roots in the independent evaluators formula for  $\hat{P}_\alpha$  can only occur when the ensemble classifiers are correlated in the evaluation. It signals, with no false positives, that an observed data sketch was not produced by an ensemble of independent classifiers. Unfortunately, the converse is not true precisely because of the blind spots in the algebraic evaluator discussed in the previous section.



**Theorem 5.** The presence of unresolved square roots in the independent evaluator estimate of  $\hat{P}_\alpha$  means the classifiers are not sample independent. The converse is not true. Integer ratio estimates of  $\hat{P}_\alpha$  are possible at the algebraic evaluator blind spots.

*Proof.* An unresolved square root in the  $\hat{P}_\alpha$  estimate cannot be produced by sample independent classifiers. This follows immediately from the exact solution in Theorem 2.  $\square$

## 7 Detailed comparison with Jaffe et al. [2015]

The natural counterpart to the exact solution provided in Theorem 2 is Theorem 1 in Jaffe et al. [2015]. There are various reasons for this. As mentioned in the Previous Work section, their paper follows the spectral approach to evaluation started by Parisi et al. [2014]. As such, there are no hyperparameters as in Bayesian approaches to evaluation. But like all probability approaches for evaluation, their work is concerned with inferring unknown distributions in the limit of infinite sample sizes.

The key to understanding the difference between any probabilistic approach and the one taken here is that they are discussing different notions of independence. The distributional independence assumption in Jaffe et al. [2015] is not equivalent to the sample independence assumption used in this paper. This is illustrated by the considering how to construct data sketches of independent classifiers under either approach. In the algebraic approach taken here you could use the generating set of polynomials in Theorem 1 to do this quickly. Set the prevalence and label accuracies to desired integer ratios, plug them into the polynomials and compute the resulting data sketch. The computed decision event frequencies can then be used to compute the minimum test size that would have given that data sketch - the GCD of the frequencies. We can then just do random shuffles of these decision events to create a very large number of stream simulations, all of which have zero correlation on the test and have exactly the same label prevalences and classifier accuracies. Consider now trying to get the same data sketch by simulating an i.i.d. process using the same settings as was done with just the algebra of the generating set. The resulting data sketch will almost certainly not be one that has zero sample correlations.

Their solution to distributionally independent classifiers is based on three assumptions. They are,

1. The stream items ('instances' in their paper) are assumed to be i.i.d. realizations of the marginal for them,  $p_X(x)$ , for an unknown joint distribution of the items and the classifiers decisions.
2. The classifiers are conditionally independent for any pair - their joint decision distribution is the product of their individual distributions.
3. More than half the classifiers are assumed to have label accuracies that sum to more than 1.

Assumption 1 is irrelevant in algebraic evaluation. It does not matter if the stream items are not i.i.d. under any distribution. Sample statistics about per-item events have nothing to do with sample statistics across items. Assumption 2 is the distributional definition of independence. Assumption 3, as they themselves mention, is strictly speaking not necessary. It relates to 'decoding' the true evaluation point. The point of view of this paper is that this is a separate concern, as discussed in Section 6 of the paper. Identifying the true point will always require additional side-information. Assumption 3 is akin to error-correcting codes deciding that the least bit flips solution is the correct one when an error is detected and corrected. Relying on prevalence knowledge is another acceptable way to decode the true evaluation point.

The terminology of Jaffe et al. [2015] uses *specificity* and *sensitivity* to describe what are here called the label accuracies. Interestingly, later in the paper, when they consider 3 or more label classification, they revert to using the terminology of confusion matrices - the one essentially used here. To continue the comparison with their work, we arbitrarily decide that what they call label "1" is  $\alpha$  and their label "-1" is  $\beta$ . Under that identification. the mapping to sensitivity and specificity variables is,

$$\psi_i \rightarrow \hat{P}_{i,\alpha} \tag{47}$$

$$\eta_i \rightarrow \hat{P}_{i,\beta} \tag{48}$$

But care must be taken not to equate the two. The specificity and sensitivity are referring to unknown, discrete distributions. The label accuracies in this paper refer to sample averages, not distributions. The predictions of the classifiers are represented by functions,  $f_i(X)$ , that yield ‘1’ and ‘-1’. This entanglement of labels and values then leads them to consider the infinite sample size quantities,

$$\mu_i = \mathbb{E}[f_i(X)] \quad (49)$$

$$R = \mathbb{E}[(f_i(X) - \mu_i)(f_j(X) - \mu_j)] \quad (50)$$

The crux of the spectral approach is that the off-diagonal elements of the matrix  $R$  are identical to a rank-one matrix  $\mathbf{v}\mathbf{v}^T$  where the vector  $\mathbf{v}$  encodes the class imbalance in the labels and the average of the label accuracies. But these quantities are not known for any finite sample so they must consider their finite test averages and consequently obtain noisy estimates of the quantity  $\mathbf{v}$ . They show that if one knows the class imbalance, denoted by  $b$  in their paper, then the specificities and sensitivities have consistent estimators with errors of order  $1/\sqrt{n}$ .

Let us briefly consider the relation between the finite sample estimate of the  $\mu_i$  quantities to see their relation to the work here.

$$\hat{\mu}_i = \frac{1}{n} \sum_{s=1}^n f_i(x_s) \quad (51)$$

$$= \frac{1}{n} \sum_{\mathbf{1}_s(\alpha)=1} (2\mathbb{1}_i(\ell_i^{(s)}) - 1) + \frac{1}{n} \sum_{\mathbf{1}_s(\beta)=1} (1 - 2\mathbb{1}_i(\ell_i^{(s)})) \quad (52)$$

$$= -1 + 2\hat{P}_\alpha (\hat{P}_{i,\alpha} + \hat{P}_{i,\beta} - 1) \quad (53)$$

Note that this reformulation makes clear why the condition that the label accuracies must sum to a value greater than one is not just to decode the true point. As in the work here, the spectral method approach also has blind spots. The spectral blind spot is a line,

$$\hat{P}_{i,\alpha} + \hat{P}_{i,\beta} = 1. \quad (54)$$

Finally, their solution for independent classifiers then finishes with their Theorem 1, a proof that a restricted likelihood estimator for the class imbalance will converge to its true value in probability as  $n \rightarrow \infty$ . It is tempting to conclude that the independent evaluator is much better because it provides sharp estimates that are exact for a finite sample. But this is not a fair comparison. As was remarked in the section discussing the unresolved square root in the estimate of  $\hat{P}_\alpha$  by the independent evaluator, the experiments carried here never observed a sample independent test. A fairer comparison between the two methods would be one that considered the error in the independent estimator for nearly sample independent ensembles versus the one provided by the spectral method for test sizes where classifiers independent in distribution would produce similar sample correlations. That comparison is not done here.

## 8 General comments on the experiments

All the code and data for the experiments is provided in accompanying files. The experiments were carried out on a MacBook Pro with an M1 chip. To greatly simplify the comparison and to mimic artificial constraints that may exist in production or field deployment the arbitrary decision was made that all the classifiers would use a logistic regression method with pinned settings as provided in the Mathematica implementations of the algorithm under their general function `Classify`. Similarly, the feature partitions used for all the experiments were exactly the same - each classifier used just three features completely disjoint from those used by the other ensemble members. In other words, no attempt was made to show that the experimental protocol is optimal or the best possible choice for any of the datasets.

The timing for all the runs was dominated by training of the classifiers. Algebraic evaluation is essentially immediate - one of its great benefits. The time taken to train 300 different feature partitions 10 times each over 10 different test sizes was about six hours.

## 8.1 The distance to the containing variety

The hypothesis that the independent evaluator estimate has non-zero distance to the containing variety is only explored empirically in this paper. Given the small value of the distances that are computed, one may wonder if this is due to floating point errors. Mathematica has the ability to compute this distance exactly when the surface is defined using integer ratios and the independent estimate is given as an algebraic number. These exact calculations take considerably longer than those done when floating point numbers are used. A few of these calculations were carried out to confirm the correctness of the floating point estimates.

## 8.2 Some evaluations

It is instructive to consider the actual evaluations that were carried out using the feature partitions in the experiments. The best evaluations were associated with the `twonorm` experiments. Something that the plots of algebraic failures would lead us to expect. If features are correlated, training protocols can only go so far in creating independent ensembles. The `twonorm` dataset is synthetic and its features were independently generated. The result of a single evaluation is shown in Figure 1. The evaluation was selected by performing 10 training/evaluation runs on the 100 feature partitions investigated in the 2nd set of experiments. Those feature partitions that succeeded in giving seemingly correct values (all of them in the `twonorm`) for the 10 runs were then searched for the one closest to the containing variety for its evaluation. The next best results for the minimum distance evaluation were obtained in the `mushroom` dataset and are shown in Figure 2. And finally, the `adult` evaluation was the worst.

If one is concerned with mitigating the principal/agent monitoring paradox, exhaustive searches of evaluating ensembles are hardly practical and are, therefore, not the focus of this paper. Much more practical when one is concerned about safe or profitable deployments is being able to handle evaluations where the classifiers are nearly independent. The algebraic approach presented here is a step in that direction but work remains on handling these cases. Knowing evaluations failed is hardly useful when that is the common case.

## 9 References

### References

- D. Cox, J. Little, and D. O’Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, 4th edition, 2015.
- Ariel Jaffe, Boaz Nadler, and Yuval Kluger. Estimating the accuracies of multiple classifiers without labeled data. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 407–415, San Diego, California, USA, 2015. PMLR.
- Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

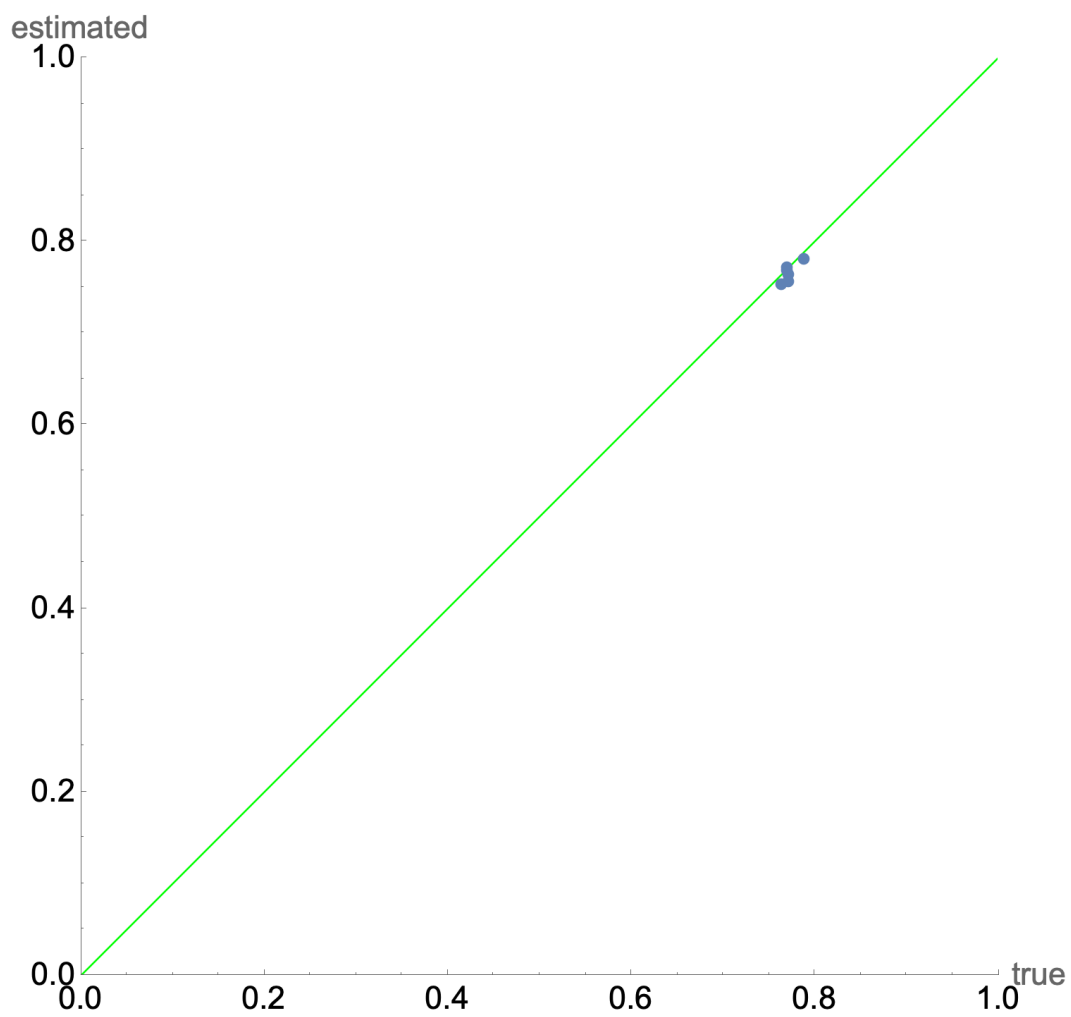


Figure 2: Estimated versus true values for the six label accuracies of three classifiers in a single evaluation of the `twonorm` features partition that had the closest distance to the containing variety.

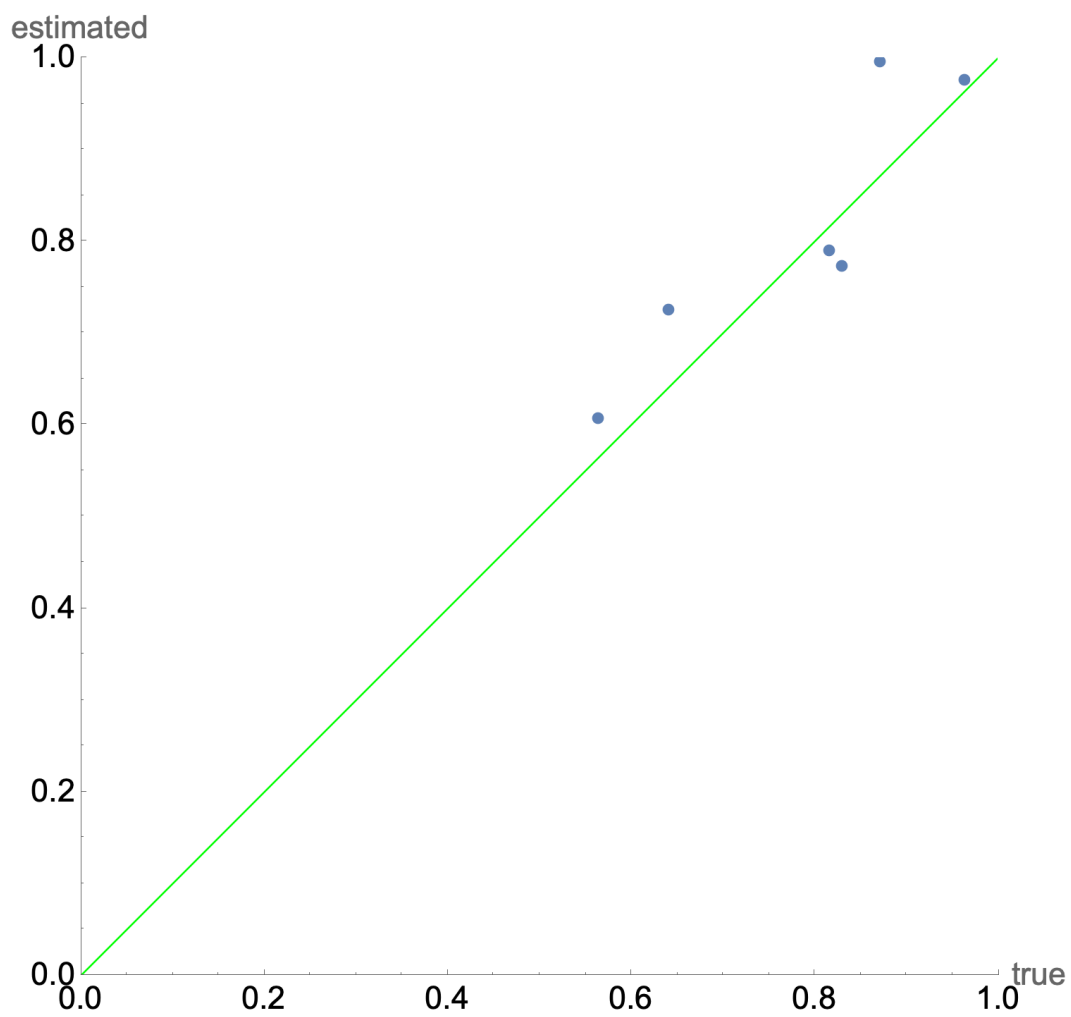


Figure 3: Estimated versus true values for the six label accuracies of three classifiers in a single evaluation of the mushroom features partition that had the closest distance to the containing variety.

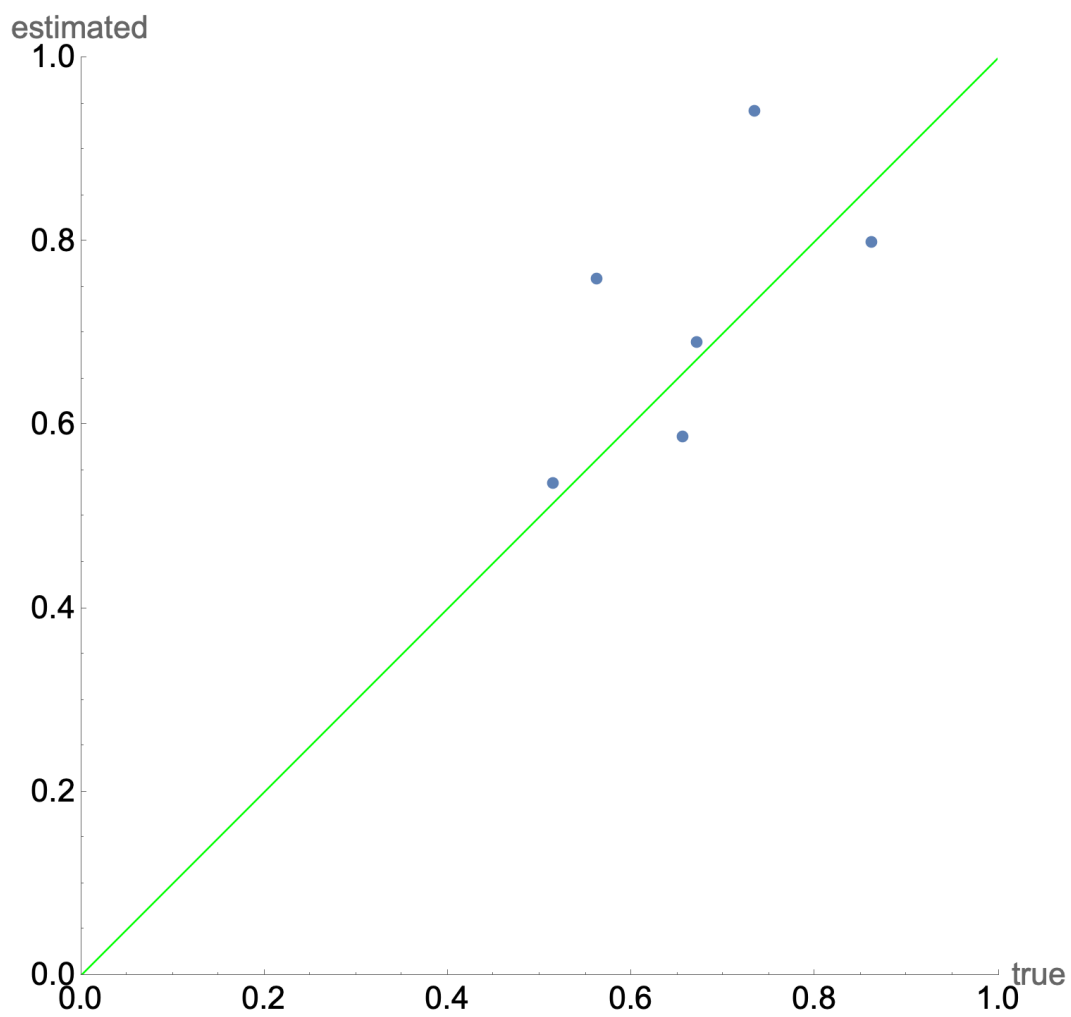


Figure 4: Estimated versus true values for the six label accuracies of three classifiers in a single evaluation of the mushroom features partition that had the closest distance to the containing variety.