Ground truth inference of binary classifier accuracies - the independent classifiers case

The accuracy of binary classifiers can be estimated exactly without knowing the true labels for the classifier decisions. Three independent classifiers are sufficient. This notebook demonstrates the above claims by construction.

The method proposed here is algebraic as opposed to the maximum-likelihood methods commonly seen in ground truth inference algorithms. This approach has been patented by Data Engines Corporation.

The Polynomial System - Algebra of Error for Independent Binary Classifiers

The following set of quartic polynomials define algebraic system that must be solved to estimate the accuracy of three classifiers as well as the true prevalance of the labels. There are 8 = 2^3 polynomials, one for each of the possible voting pattern of three classifiers doing binary label classification.

In[457]:=

$$
\begin{aligned}
&\texttt{groundTruthPolynomials} = \{-f_{\alpha,\alpha,\alpha} + P_\alpha P_{1,\alpha,\alpha} P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\alpha,\beta} P_{3,\alpha,\beta},\\
&\quad -f_{\alpha,\alpha,\beta} + P_\alpha P_{1,\alpha,\alpha} P_{2,\alpha,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\alpha,\beta} P_{3,\beta,\beta},\\
&\quad -f_{\alpha,\beta,\alpha} + P_\alpha P_{1,\alpha,\alpha} P_{2,\beta,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\beta,\beta} P_{3,\alpha,\beta},\\
&\quad -f_{\alpha,\beta,\beta} + P_\alpha P_{1,\alpha,\alpha} P_{2,\beta,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\beta,\beta} P_{3,\beta,\beta},\\
&\quad -f_{\beta,\alpha,\alpha} + P_\alpha P_{1,\beta,\alpha} P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\alpha,\beta} P_{3,\alpha,\beta},\\
&\quad -f_{\beta,\alpha,\beta} + P_\alpha P_{1,\beta,\alpha} P_{2,\alpha,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\alpha,\beta} P_{3,\beta,\beta},\\
&\quad -f_{\beta,\beta,\alpha} + P_\alpha P_{1,\beta,\alpha} P_{2,\beta,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\beta,\beta} P_{3,\alpha,\beta},\\
&\quad -f_{\beta,\beta,\beta} + P_\alpha P_{1,\beta,\alpha} P_{2,\beta,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\beta,\beta} P_{3,\beta,\beta}\}
\end{aligned}
$$

Out[457]=

$$
\begin{aligned}
&\{-f_{\alpha,\alpha,\alpha} + P_\alpha P_{1,\alpha,\alpha} P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\alpha,\beta} P_{3,\alpha,\beta},\\
&\ -f_{\alpha,\alpha,\beta} + P_\alpha P_{1,\alpha,\alpha} P_{2,\alpha,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\alpha,\beta} P_{3,\beta,\beta},\\
&\ -f_{\alpha,\beta,\alpha} + P_\alpha P_{1,\alpha,\alpha} P_{2,\beta,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\beta,\beta} P_{3,\alpha,\beta},\\
&\ -f_{\alpha,\beta,\beta} + P_\alpha P_{1,\alpha,\alpha} P_{2,\beta,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\alpha,\beta} P_{2,\beta,\beta} P_{3,\beta,\beta},\\
&\ -f_{\beta,\alpha,\alpha} + P_\alpha P_{1,\beta,\alpha} P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\alpha,\beta} P_{3,\alpha,\beta},\\
&\ -f_{\beta,\alpha,\beta} + P_\alpha P_{1,\beta,\alpha} P_{2,\alpha,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\alpha,\beta} P_{3,\beta,\beta},\\
&\ -f_{\beta,\beta,\alpha} + P_\alpha P_{1,\beta,\alpha} P_{2,\beta,\alpha} P_{3,\alpha,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\beta,\beta} P_{3,\alpha,\beta},\\
&\ -f_{\beta,\beta,\beta} + P_\alpha P_{1,\beta,\alpha} P_{2,\beta,\alpha} P_{3,\beta,\alpha} + P_\beta P_{1,\beta,\beta} P_{2,\beta,\beta} P_{3,\beta,\beta}\}
\end{aligned}
$$

The Ground Truth and the Inference Problem

In classification problems, the ground truth is the correct label for each data point in a dataset. The statistics of ground truth solved by the above polynomial system are:
- the prevalence of the labels.
- the accuracies of the classifiers for each label.

Synthetic/Fake/Manufactured Ground Truth

The purpose of this notebook is to demonstrate the mathematics of ground truth inference with polynomials. So to keep it simple, we will be manufacturing our own ground truth. The following commands create the ground truth for this problem - a list of labels for the true label of each data point in the dataset.

```
In[403]:= desiredALabelPrevalence = 23 / 100
          datasetSize = 10 000
          labelGroundTruth =
            Table[If[RandomReal[] < desiredALabelPrevalence, α, β], {datasetSize}];
           23
Out[403]= ───
          100

Out[404]= 10 000
```

Let’s check that the ground truth makes sense

```
In[407]:= RandomSample[labelGroundTruth, 10]

Out[407]= {α, β, α, β, β, β, β, α, β, β}
```

```
In[408]:= labelGroundTruth // Tally

Out[408]= {{β, 7675}, {α, 2325}}
```

Now we construct the synthetic data for the classifier label decisions. Again, we are using synthetic data to create a simple example that focuses on the math of the algorithm. You can substitute numbers from real independent classifiers if you have them.

1. The first thing we need to specify is the accuracy of each of the classifiers. It is this accuracy that we seek to estimate with the polynomial system that is constructed from the classifier decisions alone. I’ll arbitrarily pick them uniformly from the same range. I need two accuracies, one for each label since we are doing binary classification.

```
In[409]:= classifierAccuracies = Table[
            {(* The α accuracy *) α → RandomReal[{0.7, 0.9}],
              (* The β accuracy *) β → RandomReal[{0.7, 0.9}]} // Association, {3}]

Out[409]= {⟨|α → 0.724074, β → 0.762341|⟩,
           ⟨|α → 0.873416, β → 0.777599|⟩, ⟨|α → 0.758298, β → 0.757377|⟩}
```

2. Now we need a function that takes a classifier’s accuracies, a list true labels, and produces a sample of its classification decisions.
- OtherLabel is a convenience function to create incorrect classification decisions.
- SyntheticClassification is the actual function that produces a sample of what a classifier with the specified accuracies would produce on a dataset.

```
In[410]:= OtherLabel[α] = β
          OtherLabel[β] = α

Out[410]= β

Out[411]= α
```

```
In[412]:= SyntheticClassification[classifierAccuracies_Association, dataset_List] :=
           dataset //
            Map[
              If[
                (* Throw a die, if below accuracy,
                pick the true label, if not, pick the other label *)
                RandomReal[] < classifierAccuracies[#], #, OtherLabel[#]] &,
              #] &
```

3. The classifier decisions are next. Three lists, one for each classifier, as long as the dataset.

```
In[413]:= classifierDecisions =
            Map[SyntheticClassification[#, labelGroundTruth] &, classifierAccuracies];
```

Let’s make sure things are okay

```
In[416]:= classifierDecisions[[1]] // {RandomSample[#, 10], # // Length} &
          classifierDecisions[[2]] // {RandomSample[#, 10], # // Length} &
          classifierDecisions[[3]] // {RandomSample[#, 10], # // Length} &

Out[416]= {{α, β, β, β, β, α, α, β, α, β}, 10 000}

Out[417]= {{β, β, β, β, β, β, α, β, α, α}, 10 000}

Out[418]= {{β, α, β, β, β, α, β, β, β, β}, 10 000}
```

Let’s also make sure that we are getting noisy classification of the true labels

```
In[419]:= Map[Tally, classifierDecisions]

Out[419]= {{{β, 6385}, {α, 3615}}, {{β, 6287}, {α, 3713}}, {{α, 3653}, {β, 6347}}}
```

4. Now we can construct the left side input of the ground truth polynomials for independent binary classifiers - the frequency of the voting patters

```
In[423]:= votingPatternCounts = classifierDecisions // Transpose // Tally

Out[423]= {{{β, β, α}, 1181}, {{β, β, β}, 3433}, {{α, α, β}, 695}, {{α, β, α}, 526},
           {{β, α, β}, 1072}, {{α, β, β}, 1147}, {{α, α, α}, 1247}, {{β, α, α}, 699}}
```

Let’s sort them and pretty print them to see the pattern better

```
In[426]:= votingPatternCounts // Sort // Column

          {{α, α, α}, 1247}
          {{α, α, β}, 695}
          {{α, β, α}, 526}
Out[426]= {{α, β, β}, 1147}
          {{β, α, α}, 699}
          {{β, α, β}, 1072}
          {{β, β, α}, 1181}
          {{β, β, β}, 3433}
```

Everything looks okay with our synthetic data. Note that votingPatternCounts no longer contains any information about the true label for any of the data points in labelGroundTruth. Hence, any algorithm that only uses votingPatternCounts - the observed frequency of classifier voting patterns, would be carrying out ground truth inference. The goal of the next section is to estimate classifierAccuracies and the prevalence of the labels using only votingPatternCounts.

Exact Solution for Three Independent Binary Classifiers

The above polynomial can be solved exactly for the unknown ground truth statistics given the tally of voting patters by the three classifiers.

The solution for the prevalence is a quadratic equation for the prevalence of, say, label A and complicated coefficients based on the 8 voting pattern frequencies of the three classifiers over the dataset. This means that there are two solutions for prevalence - x% or (1-x)%. Likewise for the accuracies of the classifiers, although there there is a symmetry between the accuracies for both labels.

This illustrates, in the case of classification tasks, the difference between accuracy and precision. Curating a very small subset of the target dataset immediately establishes the correct solution. Accuracy is cheap, precision is expensive.

In general, these polynomial systems are solved using ideas from algebraic geometry such as a Groebner basis and Buchberger’s algorithm.

Choosing an independent variables set

There are two labels so we have two prevalences that are unknown. But, by definition, they must sum up to 100% in the dataset. So one variable is enough. It is an arbitrary choice, I picked P_α.

Similar thing happens with the accuracies of the classifiers. There are two labels and they can give two answers for each label so each classifier has 4 variables. And similar to the case of the prevalence, there are two equations the 4 variables have to obey. I chose representing everything in terms of the probability of being correct given label.

If you put together the two paragraphs above, you can figure out that N independent classifiers would require 2*N + 1 variables to be solved for. So three classifiers should yield a polynomial system that can be reduced to 7 unknown variables (2*3 + 1).

I’ve encapsulated the arbitrary choice of what those 7 variables need to be in the following rules

In[456]:=
```
reduceToIndependentVariablesRules = {
   (* The prevalence *)
   P_{β} → (1 - P_{α}),
   (* The alpha label rules for each of the three classifiers *)
   (*Sequence@@Flatten@Table[P_{i,β,α}→(1-P_{i,α,α}),{i,3}]*)
   P_{i_,β,α} → (1 - P_{i,α,α}),
   (* The beta label rules for each of the three classifiers *)
   P_{i_,α,β} → (1 - P_{i,β,β})}
```

Out[456]= $\{P_{\beta} \to 1 - P_{\alpha},\ P_{i\_,\beta,\alpha} \to 1 - P_{i,\alpha,\alpha},\ P_{i\_,\alpha,\beta} \to 1 - P_{i,\beta,\beta}\}$

And let’s transform our polynomials into using an independent set of variables

In[458]:=
```
groundTruthPolynomials /. reduceToIndependentVariablesRules
```

Out[458]=

$$\begin{aligned}\{&-f_{\alpha,\alpha,\alpha} + P_{\alpha} P_{1,\alpha,\alpha} P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + (1 - P_{\alpha})(1 - P_{1,\beta,\beta})(1 - P_{2,\beta,\beta})(1 - P_{3,\beta,\beta}),\\ &-f_{\alpha,\alpha,\beta} + P_{\alpha} P_{1,\alpha,\alpha} P_{2,\alpha,\alpha}(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})(1 - P_{1,\beta,\beta})(1 - P_{2,\beta,\beta}) P_{3,\beta,\beta},\\ &-f_{\alpha,\beta,\alpha} + P_{\alpha} P_{1,\alpha,\alpha}(1 - P_{2,\alpha,\alpha}) P_{3,\alpha,\alpha} + (1 - P_{\alpha})(1 - P_{1,\beta,\beta}) P_{2,\beta,\beta}(1 - P_{3,\beta,\beta}),\\ &-f_{\alpha,\beta,\beta} + P_{\alpha} P_{1,\alpha,\alpha}(1 - P_{2,\alpha,\alpha})(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})(1 - P_{1,\beta,\beta}) P_{2,\beta,\beta} P_{3,\beta,\beta},\\ &-f_{\beta,\alpha,\alpha} + P_{\alpha}(1 - P_{1,\alpha,\alpha}) P_{2,\alpha,\alpha} P_{3,\alpha,\alpha} + (1 - P_{\alpha}) P_{1,\beta,\beta}(1 - P_{2,\beta,\beta})(1 - P_{3,\beta,\beta}),\\ &-f_{\beta,\alpha,\beta} + P_{\alpha}(1 - P_{1,\alpha,\alpha}) P_{2,\alpha,\alpha}(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha}) P_{1,\beta,\beta}(1 - P_{2,\beta,\beta}) P_{3,\beta,\beta},\\ &-f_{\beta,\beta,\alpha} + P_{\alpha}(1 - P_{1,\alpha,\alpha})(1 - P_{2,\alpha,\alpha}) P_{3,\alpha,\alpha} + (1 - P_{\alpha}) P_{1,\beta,\beta} P_{2,\beta,\beta}(1 - P_{3,\beta,\beta}),\\ &-f_{\beta,\beta,\beta} + P_{\alpha}(1 - P_{1,\alpha,\alpha})(1 - P_{2,\alpha,\alpha})(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha}) P_{1,\beta,\beta} P_{2,\beta,\beta} P_{3,\beta,\beta}\}\end{aligned}$$

And let’s check that we only have seven non-frequency variables

In[461]:=
```
groundTruthPolynomials /. reduceToIndependentVariablesRules // Variables /@ # & //
    Flatten // DeleteDuplicates // SortBy[#, First] &
```

Out[461]=

$$\{f_{\alpha,\alpha,\alpha},\ f_{\alpha,\alpha,\beta},\ f_{\alpha,\beta,\alpha},\ f_{\alpha,\beta,\beta},\ f_{\beta,\alpha,\alpha},\ f_{\beta,\alpha,\beta},\ f_{\beta,\beta,\alpha},\ f_{\beta,\beta,\beta},\ P_{\alpha},\ P_{1,\alpha,\alpha},\ P_{1,\beta,\beta},\ P_{2,\alpha,\alpha},\ P_{2,\beta,\beta},\ P_{3,\alpha,\alpha},\ P_{3,\beta,\beta}\}$$

Preparing the observed voting frequencies rules

Recall that we have the voting pattern counts already

In[462]:=
```
votingPatternCounts // Sort // Column
```

Out[462]=
```
{{α, α, α}, 1247}
{{α, α, β}, 695}
{{α, β, α}, 526}
{{α, β, β}, 1147}
{{β, α, α}, 699}
{{β, α, β}, 1072}
{{β, β, α}, 1181}
{{β, β, β}, 3433}
```

Let’s turn them into rules so we can substitute their value into the ground truth polynomials. We need to normalize them so we need the total count first. We now it is 10K, but let’s get it by computation so we can check everything is okay

```
In[465]:= totalVotes = votingPatternCounts // Last /@ # & // Total
          totalVotes == datasetSize
```

```
Out[465]= 10 000

Out[466]= True
```

```
In[467]:= frequencyRules = votingPatternCounts //
            Map[(Subscript[f, Sequence @@ #[[1]]] → #[[2]] / datasetSize) &, #] &
```

Out[467]=

$$\left\{f_{\beta,\beta,\alpha} \to \frac{1181}{10\,000},\ f_{\beta,\beta,\beta} \to \frac{3433}{10\,000},\ f_{\alpha,\alpha,\beta} \to \frac{139}{2000},\ f_{\alpha,\beta,\alpha} \to \frac{263}{5000},\ f_{\beta,\alpha,\beta} \to \frac{67}{625},\ f_{\alpha,\beta,\beta} \to \frac{1147}{10\,000},\ f_{\alpha,\alpha,\alpha} \to \frac{1247}{10\,000},\ f_{\beta,\alpha,\alpha} \to \frac{699}{10\,000}\right\}$$

The final polynomial set

```
In[468]:= finalPolynomialSet =
           (groundTruthPolynomials /. reduceToIndependentVariablesRules) /. frequencyRules
```

Out[468]=

$$\begin{aligned}\Big\{&-\frac{1247}{10\,000} + P_{\alpha}\, P_{1,\alpha,\alpha}\, P_{2,\alpha,\alpha}\, P_{3,\alpha,\alpha} + (1 - P_{\alpha})\,(1 - P_{1,\beta,\beta})\,(1 - P_{2,\beta,\beta})\,(1 - P_{3,\beta,\beta}),\\
&-\frac{139}{2000} + P_{\alpha}\, P_{1,\alpha,\alpha}\, P_{2,\alpha,\alpha}\,(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})\,(1 - P_{1,\beta,\beta})\,(1 - P_{2,\beta,\beta})\, P_{3,\beta,\beta},\\
&-\frac{263}{5000} + P_{\alpha}\, P_{1,\alpha,\alpha}\,(1 - P_{2,\alpha,\alpha})\, P_{3,\alpha,\alpha} + (1 - P_{\alpha})\,(1 - P_{1,\beta,\beta})\, P_{2,\beta,\beta}\,(1 - P_{3,\beta,\beta}),\\
&-\frac{1147}{10\,000} + P_{\alpha}\, P_{1,\alpha,\alpha}\,(1 - P_{2,\alpha,\alpha})\,(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})\,(1 - P_{1,\beta,\beta})\, P_{2,\beta,\beta}\, P_{3,\beta,\beta},\\
&-\frac{699}{10\,000} + P_{\alpha}\,(1 - P_{1,\alpha,\alpha})\, P_{2,\alpha,\alpha}\, P_{3,\alpha,\alpha} + (1 - P_{\alpha})\, P_{1,\beta,\beta}\,(1 - P_{2,\beta,\beta})\,(1 - P_{3,\beta,\beta}),\\
&-\frac{67}{625} + P_{\alpha}\,(1 - P_{1,\alpha,\alpha})\, P_{2,\alpha,\alpha}\,(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})\, P_{1,\beta,\beta}\,(1 - P_{2,\beta,\beta})\, P_{3,\beta,\beta},\\
&-\frac{1181}{10\,000} + P_{\alpha}\,(1 - P_{1,\alpha,\alpha})\,(1 - P_{2,\alpha,\alpha})\, P_{3,\alpha,\alpha} + (1 - P_{\alpha})\, P_{1,\beta,\beta}\, P_{2,\beta,\beta}\,(1 - P_{3,\beta,\beta}),\\
&-\frac{3433}{10\,000} + P_{\alpha}\,(1 - P_{1,\alpha,\alpha})\,(1 - P_{2,\alpha,\alpha})\,(1 - P_{3,\alpha,\alpha}) + (1 - P_{\alpha})\, P_{1,\beta,\beta}\, P_{2,\beta,\beta}\, P_{3,\beta,\beta}\Big\}\end{aligned}$$

Let’s check that only the seven unknown statistics of the ground truth remain in the final polynomial system.

```
In[470]:= statisticsOfGroundTruth =
           finalPolynomialSet // Variables /@ # & // Flatten // DeleteDuplicates
```

Out[470]= $\{P_{\alpha},\ P_{1,\alpha,\alpha},\ P_{2,\alpha,\alpha},\ P_{3,\alpha,\alpha},\ P_{1,\beta,\beta},\ P_{2,\beta,\beta},\ P_{3,\beta,\beta}\}$

The “magical” solution

```
In[472]:= Solve[Map[# == 0 &, finalPolynomialSet], statisticsOfGroundTruth] // N

Out[472]= {{P_α → 0.237026, P_{1,α,α} → 0.746384, P_{2,α,α} → 0.872898,
            P_{3,α,α} → 0.743696, P_{1,β,β} → 0.758068, P_{2,β,β} → 0.784526, P_{3,β,β} → 0.752253},
           {P_α → 0.762974, P_{1,α,α} → 0.241932, P_{2,α,α} → 0.215474, P_{3,α,α} → 0.247747,
            P_{1,β,β} → 0.253616, P_{2,β,β} → 0.127102, P_{3,β,β} → 0.256304}}
```

Recall what the “real” ground truth is in this manufactured case

```
In[473]:= classifierAccuracies

Out[473]= {<|α → 0.724074, β → 0.762341|>,
           <|α → 0.873416, β → 0.777599|>, <|α → 0.758298, β → 0.757377|>}
```

But note that classifierAccuracies are the dice we three to produce the dataset. So we should really be comparing our answer to the sample statistic - ground truth inference is always about the sample statistics as befits a non-parametric method!

Let’s calculate the sample accuracies of classifier 1 to see how far off the 0.746 answer is from it. We align our ground truth with the classifiers decisions

```
In[475]:= Length@classifierDecisions[[1]]

Out[475]= 10 000

In[476]:= Length@labelGroundTruth

Out[476]= 10 000

In[478]:= Transpose[{labelGroundTruth, classifierDecisions[[1]]}] // GroupBy[#, First] & //
            Map[Last, #, {2}] & // Tally /@ # &

Out[478]= <|β → {{β, 5796}, {α, 1879}}, α → {{β, 589}, {α, 1736}}|>
```

So the sample accuracy for the alpha data points for classifier 1 is

```
In[479]:= 1736 / (1736 + 589) // N

Out[479]= 0.746667
```

Ha ha! Cool. This simple experiment should clarify what it means to say that ground truth inference algorithms measure the statistic of the ground truth. And ground truth is always referring to the sample at hand in your study. The method estimated the sample statistic of classifier 1 on the alpha label - 0.746667 as 0.746384. That estimate is not about the die that I used to produce the synthetic data, the process statistic (0.724074) that is in classifierAccuracies.