Sample statistics space for self-assessment algorithms without ground truth is finite

A huge barrier to the progress of science is our ignorance of the size of the “true” model’s parameter space. Thus, no one can tell immediately if our proposed explanation for a phenomena is too complicated by having too many variables, or too simplistic by not having enough.

The space of variables needed to explain a finite sample in a data stream is not like that at all. It is both finite-dimensional and we can compute its size immediately by enumerating all statistics that we would need to exhaustively describe the statistics you want on a given portion of the data stream.

We demonstrate this claim by trying to write the EXACT polynomial system that describes all possible decision frequencies for an ensemble of arbitrarily correlated binary classifiers. This specificity is essential to the argument. There are gazillions of sample statistics that may be of interest to the user. For each set they define, a different polynomial system would be created and it would have a different dimension. But - whatever that value is - it is finite and allows for an exact polynomial description.

Note the Faustian bargain here. We gain a huge “theoretical” advantage in describing the metrology of some unknown ground truth statistics. But our metrological labor is then multiplied by the fact that many such sample statistics are possible for any sample of the data stream of a natural phenomena. This is, understandably, not attractive to scientists. But to the engineer or industrial scientist it is because it allows some clarity when we seek a few ground truth statistics so we can monitor AI systems in production or robots we want to build.

An example from our work in Speech Recognition -I was a member of the research team that helped produce the 1st commercial continuous speech recognition, Dragon Naturally Speaking - elucidates this industrial mathematician/scientist insight. Developing Naturally Speaking or when people report on speech recognizers in the literature one ground truth statistic is used above all others - Word Error Rate (WER). WER is a sample statistic - your average recognition error over the speech sample you transcribe with your recognizer.

Note that WER is not exhaustive as a descriptor of all possible sample statistics related to the performance of a speech recognizer over a finite sample of speech. We could get more detailed knowledge of the error modes of the recognizer by asking more detailed error statistics on the sample. How many

times was “foo” correctly recognized? Or “bar”? How many times was “foo” transcribed as “bar”? And so on.

The crucial point for the industrial mathematician/scientist - nobody cares about that detailed error knowledge in your industry. All research decisions, all management decisions in any particular technology business are usually driven by a very small set of sample statistics. In Speech Recognition WER is that statistic.

The Two types of sample statistics in GTI

The size of the sample statistics space for monitoring binary classifiers

Now suppose that you are an industrial scientist that has deployed a bank of binary classifiers to be used as one small component of an industrial data pipeline. How do you monitor the quality of their decisions when you don’t know the true labels for their decisions? You don’t have the true labels because if you did - why did you deploy noisy classifiers to get them?

This is a crucial business insight for an industrial scientist. Companies today may have tons of data, but ground truth on that data is the scarcest commodity! Why? Because if it wasn’t, why is your company running AI algorithms over the data? The most important ground truth is missing - the one relevant to the business bottom line. This makes ground truth inference algorithms on that unknown ground truth very valuable in the very narrow context of your business goals. It is not pretty science, but it is valuable engineering.

The advantages and disadvantages of taking a black-box approach while monitoring AI or human algorithms.

When I first introduced the notion of calculating a ground truth statistic to scientists and engineers at Nuance around 2008, one engineers criticism was very telling and is actually echoed in much of the self-assessment approaches used today. The approach you see here for monitoring is the dumbest approach possible. We will treat all the binary classifiers as black-boxes that merely produce noisy labels.

This is not the only approach one can take. You could hook into the internals of the classifiers and study their internal state to detect when you think they are in an errorful state. This characterizes most self-assessment work published today - build some detailed algorithm that exploits your knowledge of the way the classifier/AI algorithm does decisions. This works but it has many costs:
- The development of the scientific basis for using internal states of your judges to determine the truthiness of their decisions.

- The inability to generalize your approach to the internal states of other, widely different, classifiers.
- The additional computational burden of memory and time to compute the wanted ground truth statistic.

Some businesses decide to accept that cost. That explains NNs research into self-assessment by FacebookAI etc.

Our black box approach is poorer in measuring that ground truth statistic (it may require more samples than the internal state approach) but has the advantage of being more universal. For example, it would allow the inclusion of classifiers for which you do not have or want an internal state representation (a thorny scientific issue!). One important class of such classifiers - humans. In speech recognition the golden standard is human transcriptions. But what happens if the transcriptions have errors? Our approach can answer that question. Internal representation approaches cannot. This is also our critique

The Two types of sample statistics in our GTI problem

We will be concerned with two types of sample statistics in the data stream of binary classifiers decisions - the observables and the ground truth statistics.

Observables are sample statistics that do not require any knowledge of the ground truth for the decisions. In this case, one observable sample statistic is the percentage of times two classifiers voted (α, β). For two binary classifiers the event space is then the set of all possible decision patterns - there are four of these, 2^2.

Ground truth statistics are sample statistics that require some knowledge of the unknown ground truth for the decisions. In the case of the binary classifiers, we want to know what percentage of the (α, β) decisions corresponded to the true label being α or β. This would allow us to calculate the accuracy on each label.

Two correlated binary classifiers

The sample statistics space for the decisions of two binary classifiers is then encapsulated in this exact polynomial description of the observables (the “f”s) in terms of the ground truth statistics (the “ϕ”s) and
and the correlations (the “Γ”s between them.

In[292]:=
```
exactPolynomialDescription = {f_{α,α} - ϕ_α (Γ_{1,2,α} + ϕ_{1,α,α} ϕ_{2,α,α}) - ϕ_β (Γ_{1,2,β} + ϕ_{1,α,β} ϕ_{2,α,β}),
  f_{α,β} - ϕ_α (-Γ_{1,2,α} + ϕ_{1,α,α} ϕ_{2,β,α}) - ϕ_β (-Γ_{1,2,β} + ϕ_{1,α,β} ϕ_{2,β,β}),
  f_{β,α} - ϕ_α (-Γ_{1,2,α} + ϕ_{1,β,α} ϕ_{2,α,α}) - ϕ_β (-Γ_{1,2,β} + ϕ_{1,β,β} ϕ_{2,α,β}),
  f_{β,β} - ϕ_α (Γ_{1,2,α} + ϕ_{1,β,α} ϕ_{2,β,α}) - ϕ_β (Γ_{1,2,β} + ϕ_{1,β,β} ϕ_{2,β,β)}}
```

Out[292]=
```
{f_{α,α} - ϕ_α (Γ_{1,2,α} + ϕ_{1,α,α} ϕ_{2,α,α}) - ϕ_β (Γ_{1,2,β} + ϕ_{1,α,β} ϕ_{2,α,β}),
 f_{α,β} - ϕ_α (-Γ_{1,2,α} + ϕ_{1,α,α} ϕ_{2,β,α}) - ϕ_β (-Γ_{1,2,β} + ϕ_{1,α,β} ϕ_{2,β,β}),
 f_{β,α} - ϕ_α (-Γ_{1,2,α} + ϕ_{1,β,α} ϕ_{2,α,α}) - ϕ_β (-Γ_{1,2,β} + ϕ_{1,β,β} ϕ_{2,α,β}),
 f_{β,β} - ϕ_α (Γ_{1,2,α} + ϕ_{1,β,α} ϕ_{2,β,α}) - ϕ_β (Γ_{1,2,β} + ϕ_{1,β,β} ϕ_{2,β,β)}}
```

Note the “chicken-and-egg” problem of self-assessment. These polynomials do not factor (an algebraic claim that can be proven or disproven). One class of ϕ variables is completely environmental (the prevalence of the true labels in the sample being analyzed). The other class is the instrument’s marginal performance in that same sample.

This chicken-and-egg nature of self-assessment results in a much deeper realization. The boundary between experimental and environmental error may be theoretically definable by your model of the world but it can never be made infinitely thin by any experimental set-up you use to explore it. You can engineer better and better measuring protocols to make the boundary as thin as you want - but it is impossible to resolve it infinitely sharp. A small insight about the deficiency of empirical measurements to fully resolve all theoretical claims - that the boundary between the machine and the experiment exists EXACTLY where you say it is. Like all of science we can only be left with a very small uncertainty of our theoretical assertions.

The size of this sample statistic space

If we treat the observables as coefficients, we can count in this case what the dimensionality of the space is. It is 2 + 2*4 + 2 = 12. By utilizing the logical relations between some of these statistics - e.g. the prevalences must sum to one - we can reduce the dimensionality of the polynomial system to 7.