Mathematical Theory of Data Accuracy

Ask an analyst to find issues in a dataset and they'll come back in a few hours with a long list. Ask them instead to *validate* a dataset and they'll be busy for months.

> Data inaccuracy is detectable, data accuracy is not.

What we will cover

What can we *know* about the data

- **Maximal Alert Theorem** - The *only* way to find issues is to look for constraint violations
- **Quality Coverage Equation** - What percent of issues are findable
    - There are always some unfindable issues

The Setup

Let $\hat{X} = (x_1, x_2 ... x_N)$ be observable datapoints measuring some 'true' values $\hat{T} = (t_1, t_2 ... t_N)$.

For now we assume that the data is discrete and only takes integer values.

Definition: Data Accuracy Issue

A data accuracy issue is simply when $\hat{X} \neq \hat{T}$.

Note: We are only dealing with discrete data for this blog post. This definition will be amended when dealing with continuos data.

Example

Var	Definition
t_1	The *true* total cars entering the ferry at the first stop.
x_1	The *observed* total cars entering the ferry at the first stop.
t_2	The *true* total cars leaving the ferry at the second stop.
x_2	The *observed* total cars leaving the ferry at the second stop.

You can't see t_1 and t_2 directly, but we hope that our volunteer did a dutiful job so that $x_1 = t_1$ and $x_2 = t_2$.

Data Accuracy Issues at the Ferry

Let's take a look at some of the data from the ferry.

day	x_1	x_2
Monday	9	5
Tuesday	7	3
Wednesday	5	7

Now I'm going to append the true number of cars entering and exiting. We normally don't get to see this.

day	t_1	t_2	x_1	x_2
Monday	9	5	9	5
Tuesday	8	3	7	3
Wednesday	7	5	5	7

A data accuracy issue is whenever $\hat{T} \neq \hat{X}$. Let's mark the rows with data accuracy issues.

day	t_1	t_2	x_1	x_2	Is DQ Issue
Monday	9	5	9	5	No
Tuesday	**8**	3	**7**	3	Yes
Wednesday	**7**	**5**	**5**	**7**	Yes

For Tuesday $t_1 \neq x_1$ and Wednesday has $t_1 \neq x_1$ and $t_2 \neq x_2$ so these two days have data accuracy issues.

Adding Real World Assumptions

We introduce an assumption g,

- $g(\hat{T}) = \mathrm{True}$ if the assumption holds
- $g(\hat{T}) = \mathrm{False}$ if the assumption is violated.

Generally we can have many real-world assumptions: $g_1, g_2 .. g_M$.

Definition: Plausible

If $\hat{T}$ satisfies all constraints $g_1, g_2..g_M$, we call $\hat{T}$ plausible.

Real World Assumptions Example

Let's go back to the ferry example.

A safe assumption is that $t_1 \geq t_2$. In other words, we can't have more cars leave the ferry than get on the ferry.

Therefore we define our real-world assumption: $g(t_1, t_2) = t_1 \geq t_2$. Let's call this assumption the *conservation of cars* assumption.

General Data Accuracy Alerts

A data accuracy alert monitors data and fires when a data accuracy issue is detected.

Let's come up with a definition given our framework.

Definition: Deterministic Alert

An alert that fires only when there is a plausible data accuracy issue. More formally:

A binary function of the data , f, is a *deterministic alert* if:

When $\hat{T}$ is plausible and $f(\hat{X})$ has value True then $\hat{T} \neq \hat{X}$.

In other words, if a deterministic alert fires that means something is wrong with the data. It probably won't catch all of the data accuracy issues, but it should never fire when the data is fine. We don't care about how the alert behaves when the true values aren't plausible because that will never happen.

The next example will make this more concrete.

Deterministic Alert Example

Let's come up with a deterministic alert for our ferry problem.

We know that $t_1 \geq t_2$ by the conservation of cars assumption. Therefore, if we see $x_1 \leq x_2$ then we know we have a data accuracy issue.

Let's define our deterministic alert a as $a(x_1, x_2) = x_1 < x_2$. When a is True we know there is a data accuracy issue.

Note: We have not yet proven that a satisfies the formal definition of a deterministic alert. The next section contains a proposition that shows that a indeed satisfies the definition.

Given the ferry problem setup, let's try to find data accuracy issues:

day	x_1	x_2	$a(x_1, x_2)$
Monday	9	5	False
Tuesday	7	3	False
Wednesday	5	7	True

Now let's join back with the true values to see how good this alert is.

day	t_1	t_2	x_1	x_2	Is DQ Issue	$a(x_1, x_2)$
Monday	9	5	9	5	No	False
Tuesday	8	3	7	3	Yes	False
Wednesday	7	5	5	7	Yes	True

Our data accuracy alert caught the issue on Wednesday, but it did not catch the issue on Tuesday.

Maximal Alert

It is natural to ask - what is the most comprehensive alert on the data? Is there an alert that will catch every issue?

Proposition: If a_1 and a_2 are deterministic alerts, so is $(a_1 \text{ or } a_2)$.

Proof omitted.

We can imagine a *maximal deterministic alert* which fires if *any* other deterministic alert fires. If we magically knew all other deterministic alerts we could create this maximal alert by *or-ing* them all together.

Definition: Maximal Alert

The alert constructed by 'or-ing' all other deterministic alerts together. Therefore the maximal alert fires iff one or more other deterministic alerts fire.

Alert Example

Proposition: $\neg g(\hat{X})$ is a deterministic alert for any constraint g.

We need to show that if $\neg g$ fires and the setup is plausible that $\hat{T} \neq \hat{X}$.

- If the alert fires we have $g(\hat{X}) = \mathrm{False}$.
- If the setup is plausible we have $g(\hat{T}) = \mathrm{True}$.

Therefore $g(\hat{T}) \neq g(\hat{X})$. This implies that $\hat{T} \neq \hat{X}$.

For a problem with real-world assumptions $g_1, g_2 ... g_M$ we also consider the *assumption violation alert* (or ava) which fires when *any* assumption is violated. Formally we can define the *ava* as follows.

$$ava(\hat{X}) = \neg g_1(\hat{X}) \text{ or } \neg g_2(\hat{X}) ... \text{ or } \neg g_M(\hat{X})$$

Where the $\neg$ symbol means 'not'.

Proposition: The assumption violation alert is a deterministic alert.

Proof:

Now that we have shown $\neg g(\hat{X})$ is a deterministic alert, since the ava is an "or" operation on deterministic alerts, it is an alert.

Q.E.D.

Theorem: Maximal Alert Theorem: *The Maximal Alert is the Constraint Violation Alert.*

We must show that if one of these alerts fire, so does the other one.

Step 1: Maximal Alert fires when Constraint Violation Does

If the constraint violation alert fires, so does the maximal one since the maximal alert fires if any alert fires.

Step 2: Constraint Violation Fires when Maximal Alert fires

We can show this by contradiction:

Say there exists a $(T = \hat{T}^*, X = \hat{X}^*)$ where the maximal alert fires and the constraint violation alert doesn't.

Now consider $(\hat{T} = \hat{X}^*, \hat{X} = \hat{X}^*)$.

- The maximal alert will fire since $\hat{X}$ is unchanged.
- There are no data accuracy issues since $\hat{X} = \hat{X}^* = \hat{T}$.
- $\hat{T}$ does not violate any constraints because $\hat{X}$ doesn't violate ava and $\hat{X} = \hat{T}$. In other words, $\hat{T}$ is *plausible*.

The maximal alert can't fire in a plausible situation with no data accuracy issues since this contradicts the definition of a deterministic alert. Therefore the maximal alert can't fire when the constraint violation alert is silent.

Q.E.D.

Implications

> The *only* way to detect data accuracy issues is to make real-world assumptions and look for violations.

Accuracy Coverage Equation

In a data accuracy system let P denote the set of $(t_1 ... t_N)$ that are plausible (don't violate real-world assumptions). Let $|P|$ denote the size of the plausible set.

Let $|\hat{X}|$ be the number of states the data can be in. For example, in a binary system $x_1 ... x_N$ can be in 2^N states.

Theorem: Accuracy Coverage Equation

Fraction of Accuracy Issues Detected = $\frac{|\hat{X}|-|P|}{|\hat{X}|-1}$

Proof

Size state space of $(\hat{T}, \hat{X})$ where $\hat{T}$ is plausible= $|P|\,|\hat{X}|$

- State space of plausible $\hat{T}$ times state space of $\hat{X}$

States with no issues = $|P|$

- Every plausible $\hat{T}$ has only one state $\hat{X}$ without accuracy issues ($\hat{X} = \hat{T}$).

Data Accuracy Issues = $|P||\hat{X}| - |P|$

- Total system states minus states without issues

Detectable Issues = $|P|(|\hat{X}| - |P|)$

- State space of plausible $\hat{T}$ times state space of $\hat{X}$ which violates ava.

Fraction of Accuracy Issues Detected = $\frac{|P|(|\hat{X}|-|P|)}{|P||\hat{X}|-|P|} = \frac{|\hat{X}|-|P|}{|\hat{X}|-1}$

Data Issues are Detectable, Data Correctness is Not

Looking at the coverage equation $\frac{|\hat{X}|-|P|}{|\hat{X}|-1}$ we see that we get 100% coverage only in the case where $|P| = 1$, which is a trivial case. In the case that $|P| = 1$, there was no reason to gather any data in the first place, since $t_1..t_n$ is fully determined by the constraints. Therefore we can say the following:

In non-trivial discrete data systems there will always be accuracy issues that are undetectable.

Sadly, this also carries over to the continuous case, which we will cover in the future.

As we gather more expectations about the real-world, we can add constraints, which will lower $|P|$, and raise the data coverage.