ML for Sci

J Jensen
UKRI-STFC

27 June 2019

Towards a Formal Framework for Facilities Science

We examine an as far as the author knows previously unpublished formalisation of facilities experiments proposed by Prof. Jared Tanner (Oxford), *et al.*

Our aim is to examine and interpret its content, first in the context of a very simple science model (we roll a die. . .), and then later in the wider context of machine learning for sciences in general, and facilities in particular.

Slides will be made available.

ToC

- Introduction of the model of Jared Tanner *et al*
- A digression into the philosophy of science
- Finding more ML applications
- A note on model selection
- Conclusion – if we learnt something, what did we learn?
- Future directions – in the unlikely event we want to know more
- References

Introducing Tanner's Model

X

We start with X which is the set of (physical) parameters that we (intend to) measure in an experiment.

Simple example: rolling a die (once), p is the (true, physical) probability of rolling a 6.

Here, $X = [0, 1]$ because $p \in X$ is a probability.

Another way of describing the die is to introduce the probability p_i of outcome i, $i = 1 \dots 6$. In this case,

$$X = \left\{ (p_i) \in \mathbb{R}^6 \middle| \sum_i p_i = 1 \wedge \forall i : p_i \geq 0 \right\}$$

A true die would have $\forall i : p_i = \frac{1}{6}$, of course. We shall return to these models later.

Introducing Tanner's Model

$$X \xrightarrow{E} Y$$

In an actual (classical) experiment, we might wish to test the hypothesis $\mathcal{H} : p = \frac{1}{6}$ by rolling the die N times and check if (say) the number of 6es is approximately $N/6$.

Introducing Y as the *measurements* of the experiments. Say E is rolling the die 1000 times and observing $y_E \in Y$ sixes.

In this case, $Y_E = \{0, \ldots, 1000\}$.

Introducing Tanner's Model

$$\begin{tikzcd} X \arrow[r, "E"] & \mathcal{Y}_E \end{tikzcd}$$

However, the actual *measurement* of E is actually a stochastic variable $\underline{\overline{Y}}_E$ with (initially) unknown distribution. Or unknown parameters of the distribution. We introduce $\mathcal{Y}_E$ as the set of possible stochastic variables for experiment E. In our die example,

$$\mathcal{Y}_E = \left\{ \underline{\overline{Y}} \mid \underline{\overline{Y}} \sim \mathrm{Ber}(1000, x), x \in X \right\}$$

Introducing Tanner's Model

$$\begin{tikzcd}
x \arrow[r, mapsto] & \underline{\overline{Y}}_E \arrow[d, mapsto] \\
 & m
\end{tikzcd}$$

- For the (first) die example, x is the (unknown) probability of a six in a single roll.
- The x value maps to a stochastic variable $\underline{\overline{Y}}_E \sim \mathrm{Ber}(N, x)$
- The *value* of the stochastic variable is the value of the *measurement* $y \in Y$ which here is the count of sixes in E, i.e. 1000 rolls of the die.
- The value y lets us define a *model* m, in some to-be-defined model-space M.

The measurement y (and some stats theory) lets us estimate

$$\hat{x} = y/N$$

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, "E"] & \mathcal{Y}_E \arrow[d] \\
 & M
\end{tikzcd}$$

The *model* helps us *describe* the Thing that we are examining.

The two models we've encountered so far are both summarised as:

- Each roll of the die is independent of any past roll
- There are precisely six outcomes, $\{1, \ldots, 6\}$
- There is a constant probability p for getting a six.
- Each outcome i occurs with a constant probability p_i
  - Hence $\sum_{i=1}^{6} p_i = 1$

We introduce a set M of models.

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, "E"] & \mathcal{Y}_E \arrow[d] \\
& M
\end{tikzcd}$$

The *model* describes our knowledge of the die, independent of the experiment which led to this knowledge.

The model can be:

- *Validated* against experimental data (more on this later)
- *Parameterised* based on results of experiments (next slide)
- Used to make *predictions* (more on this later)

Introducing Tanner's Model

$$\begin{tikzcd} X \arrow[r, "E"] & \mathcal{Y}_E \arrow[d] \\ & M \end{tikzcd}$$

Notice that, like $\underline{\overline{Y}}_E$, the models are *parameterised*. So far, M consists of m_1 and m_2:

- Model $m_1(p)$ says the die rolls are independent with sixes appearing with constant probability p.
- Model $m_2((p_i))$ says the die rolls are independent with outcome i appearing with constant probability p_i.

Note also that y (the value of $\underline{\overline{Y}}_E$) lets us estimate the parameters (it is a *sufficient statistic*), but is not sufficient to check independence.

Note also that the models are by no means stochastic; only their parameters depend on random processes.

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, "E"] & \mathcal{Y}_E \arrow[d, "?"] \\
& M?
\end{tikzcd}$$

Notice that E and the assumption of *independence* really say

$$\underline{\overline{Y}}_E = \sum_{j=1}^{N} \underline{\overline{D}}_j$$

where $\underline{\overline{D}}_j$ are IIDRV and take the value 1 iff the die shows six, and 0 otherwise.

This means that (for a fixed experiment), the outcome y has a PDF,

$$f_Y(y) \triangleq \mathbb{P}(Y = y) = \mathbb{P}\Big(\big|\{j \in Y | \underline{\overline{D}}_j = 1\}\big| = y\Big)$$

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, "E"] & \mathcal{Y}_E \arrow[d, "?"] \\
& M?
\end{tikzcd}$$

In our case, $Y_E = \{1, \ldots, 1000\}$, as a set. However, XXX speaking, is the set of probability *measures* on Y. Geometrically, it is a simplex in 1001 dimensions:

$$\left\{ x_0, \ldots, x_{1000} \middle| \sum_i x_i = 1 \land \forall i : x_i \geq 0 \right\}$$

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, "E"] \arrow[dr, dashed] & \mathcal{Y}_E \arrow[d] \\
 & \mathcal{P}_M
\end{tikzcd}$$

Composing the two arrows give us a map from X directly to the model

- which is what we implicitly used in the initial description of the die,
- because otherwise it's kind of hard to describe how the die works.

Introducing Tanner's Model

$$\begin{tikzcd}
X \arrow[r, shift left, "E"] \arrow[dr, dashed, shift left] & \mathcal{Y}_E \arrow[l, shift left] \arrow[d] \\
 & \mathcal{P}_M \arrow[ul, shift left]
\end{tikzcd}$$

Notice also how the *estimators* give us an inverse from the sufficient statistic(s) to X.

- Except they are not necessarily (exact) inverses
- And the map $M \to X$ reflects the usefulness of the chosen model as a means of describing the Thing that we study.

Introducing Tanner's Model

$$\begin{tikzcd}
x \arrow[r, mapsto, shift left, "E"] \arrow[dr, mapsto, "\mathcal{R}_{\theta_1}" description, phantom, start anchor=north east] \arrow[dr, mapsto, "\mathcal{D}_{\theta_2}"'] & y \arrow[l, mapsto, shift left, "\mathcal{R}_{\theta_1}"] \arrow[d, mapsto] \\
 & m(\hat{x}(y))
\end{tikzcd}$$

Introduce two mappings:

- $\mathcal{R}_{\theta_1} : Y \to X$, $\theta_1 \in \Theta_1$
- $\mathcal{D}_{\theta_2} : X \to M$, $\theta_2 \in \Theta_2$

- The "reconstruction" $\mathcal{R}$ estimates x from the experimental outcome y.
- The "feature extraction and model builder" $\mathcal{D}$ formalises our derivation of the model based on experimental data.
  - Except it also includes the *choice of model*
- Loss function $L_X(\theta_1) \triangleq \mathbb{E}_{X,\mathcal{Y}}(d_X(\mathcal{R}_{\theta_1}(y), x))$
- Loss function $L_M(\theta_2, \theta_1) \triangleq \mathbb{E}_{X,M}(d_M(\mathcal{D}_{\theta_2}(x), m(\mathcal{R}_{\theta_1}(y))))$
  - But think of minimising $L_M(\theta_2, \theta_1)$ over θ_2 once the optimal θ_1 is found... (discuss!)

Introducing Tanner's Model

$$\begin{tikzcd}
x \arrow[r, mapsto, shift left, "E"] \arrow[dr, mapsto, "\mathcal{D}_{\theta_2}"'] & y \arrow[l, mapsto, shift left, "\mathcal{R}_{\theta_1}"] \\
 & m(\mathcal{R}_{\theta_1}(y))
\end{tikzcd}$$

- (We'll return to the loss function in a mo. . .)
- Instead of $\hat{x}$, use the more general $\mathcal{R}_{\theta_1}(y)$.
- For the dice, the default $\mathcal{R}$ are the estimators
  - For m_1, $\mathcal{R}_{\theta_1}(y) = y/N$
  - And $\Theta_1 = \{\spadesuit\}$
- And $\mathcal{D}_{\theta_2}(x) = m_{\theta_2}$ is the model selector, with $\Theta_2 = \{1, 2\}$
  - How it picks which model is as yet unspecified.
  - The model selection is *not random* except to the extent we let y decide the model.
  - (Compare AIC, BIC?)

Reconstruction

$$\begin{tikzcd}
x \arrow[r, mapsto, shift left, "E"] \arrow[dr, mapsto, "\mathcal{D}_{\theta_2}"'] & y \arrow[l, mapsto, shift left, "\mathcal{R}_{\theta_1}"] \arrow[d, mapsto] \\
 & m(\mathcal{R}_{\theta_1}(y))
\end{tikzcd}$$

- Let's look at Reconstruction $\mathcal{R}_{\theta_1} : Y \to X$
- In SAXS/SANS, there is no direct "estimator" of parameters
  - The parameter space is X
  - Tanner *et al.*: Use machine learning to infer $x \in X$.
  - Hyperparameters and coefficients are Θ_1

Reconstruction (interlude)

$$\begin{tikzcd}
& \mathcal{H} \arrow[r] & \mathcal{E} \arrow[dr, "E\mapsto\mathcal{Y}_E"] & & & \\
X \arrow[ur] \arrow[rrr, dashed] & & & \mathcal{Y} \arrow[r] & Y \arrow[r] & \{T, F\}
\end{tikzcd}$$

- Introduce a set of hypotheses, H, on X
- and the powerset $\mathcal{H} \triangleq \{\mathfrak{h} | \mathfrak{h} \subseteq H\}$
- Each set of hypotheses $\mathfrak{h} \subseteq H$ maps to an experiment $E \in \mathcal{E}$ testing the validity of the combined truth, i.e., $T \triangleq \bigwedge_{h\in\mathfrak{h}} h$ and $F \triangleq \neg T$
- Here, $\mathcal{Y} \triangleq \{\mathcal{Y}_E | E \in \mathcal{E}\}$
- and $Y \triangleq \dot{\bigcup}_{E\in\mathcal{E}} Y_E$ is the (disjoint) union of all the outcome spaces;

The experiment, in case you're wondering, is $y = \mathcal{Y}_E(\omega) \in Y_E$, where $\omega \in \Omega$ and Ω is the Universe (if you weren't, don't worry about it – Universes are a whole talk by itself.)

Finally, a test $\mathfrak{h}(y) \in \{T, F\}$ tests the hypotheses.

Reconstruction (interlude)

$$\begin{tikzcd}
& \mathcal{H} \arrow[r] & \mathcal{E} \arrow[dr] & & \\
X \arrow[ur] \arrow[rrr, dashed] & & & \mathcal{Y} \arrow[r] & \{T, F\}
\end{tikzcd}$$

- Digression – the Duhem Thesis [P. Duhem, 1906]
  - Not all sciences are equal, so think of physics
  - “An experiment tests multiple hypotheses”
  - “Validation means all hypotheses (together) are true”
  - “Can’t isolate an individual hypothesis when test is rejected”
- In practice...
  - The challenge to Science: if a test fails, identify the hypotheses that caused it to fail.
  - Note that Y would depend on E, as before; the outcome of experiment E is $y \in Y_E$.

The Hidden ML Application – Hypothesis Testing

- If a test fails, we should re-test with selected hypotheses removed
  - Remove as few as possible, then map to a new E which gives us.
  - Would be nice if we could use the same E ?!

Iacta Est Alea

- Example. Hypothesis is “the die is honest”
- Hidden hypotheses:
  - Every die roll is independent of every other;
  - The probabilities of outcome i is constant.
  - In other words, model m_2 is valid.
- Experiment: E : roll die 1000 times and count frequency of each outcome.
- Under the hidden assumptions, $\mathcal{Y}_E$ is multinomial with probabilities $p = (p_1, \ldots, p_6)$
- The Hypothesis tests $\forall i : p_i = \frac{1}{6}$

Iacta Est Alea

- Die produces $6, 2, 6, 2, 6, 6, 1, 6, 6, 1, 2, 1, 4, 4, 1, 4, 4, 2, 3, 5, 1, 1, 6, 1, 5, 5, 6, 5, 6, 3,$
- Die produces $6, 2, 5, 3, 4, 1, 6, 4, 5, 4, 1, 4, 6, 5, 4, 3, 5, 4, 6, 2, 1, 3, 5, 3, 6, 2, 2, 5, 5, 3,$
- Die produces $1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6,$

Clearly a die which wishes to cheat us can do so, with E as above.

To test for truly randomness, we need to examine *the whole sequence* (e.g. [NIST-IR6483], 189 statistical tests!)

In principle, this could be a case for machine learning – failing to detect patterns... (often, visualisation is used)

Whittling Hypotheses

- Example: 3-SAT (NP complete) [3SAT]
- Animal guessing game
  - $X = \{\text{fly,cat,snake, armadillo,}\dots\}$
  - h_1: Does it have fur? h_2: Does it have four legs?
  - $h_1 \wedge h_2 \Rightarrow \mathrm{is - cat}$
  - Decision tree: pick most important information first.
  - Adjust according to new information:
    - $h_1 \wedge h_2 \Rightarrow \neg \mathrm{is - dog}$ (it can't be both a cat and a dog)
    - $h_3 : \mathrm{says - meow}$ (what if the cat doesn't say anything?)
    - Which sort of takes us into the Quine Thesis. . .
- Whittling applies also to sensor data and biostuff
  - combinatorial methods towards explainable AI [NIST-AI]
  - Also found in software testing (combinatorial)

The first loss function revisited

$$L_X(\theta_1) \triangleq \mathbb{E}_{X,\mathcal{Y}}(d_X(\mathcal{R}_{\theta_1}(y), x))$$

- Assumption that X is now a probability space
  - Which means σ-algebra with probability measure (otherwise the expectation/integral would not exist)
  - Natural fit would be a prior on X?
  - But maybe the σ-algebra needs some thought
- Assumption that X has a metric d_X, so it's a metric space
  - Which means the σ-algebra has to contain the Borel sets
  - (induced by the topology)
- Expectation is $\int d_X(\mathcal{R}_{\theta_1}(y), x) f_{\underline{\overline{Y}}_E}(y) f_X(x) dxdy$
  - Where the f_X is the PDFs on X (the prior)
  - and $f_{\underline{\overline{Y}}_E}$ is the PDF of $\underline{\overline{Y}}_E : \Omega \to Y_E$)
  - – a weighted avg of the distance between (any) x and the predicted value of x from any experimental value y.
  - Compare KL-divergence which measures information in one probability distribution wrt another

Revisiting the Experiment

$$E : X \rightarrow \mathcal{Y}, E \in \mathcal{E}$$

- No *a priori* structure on $\mathcal{E}$
- No *a priori* structure on $\mathcal{Y}$
- So what is $E : X \rightarrow \mathcal{Y}$?
    - Measurable: a *family* of stochastic variables $\mathcal{Y}_E$

A Posterior on X? and Multimodal Experiments

If X has a prior, what is its posterior?

- The probability distribution of $\underline{\overline{Y}}_E$ is unrelated to the prior on X
- It arises from the probability distribution on Ω

$$f(x|y) = f(y|x)\frac{f(x)}{f(y)} \propto f(y|x)f(x)$$

In a multi-modal approach, the posterior for the first experiment would become the prior for the second.

Classically, sufficient statistics are good for this. But the prior may guide the choice of hypotheses $\mathfrak{h}$ and experiment E.

A Quick Note on Model Selection

Model selection is not random. . .

- Standard methods select models by best fit with fewest parameters
  - Akaike Information Criterion (AIC)
  - Bayes Information Criterion (BIC)
- In Tanner's approach, the posterior is based on ML (and hence the description of the sample in the model), so proposes ML methods for "fitness" (e.g. Wasserstein "distance")

A good fit of m_1 would beat a good fit of m_2 because $m1$ has 1 parameter and m_2 has 5 (well, 6, but 5 d.o.f.)

It would seem a good fit of m_2 includes a good fit of m_1 as a special case (in some handwavy way) and is a better description of the die?

However, a good fit to $m_0 : \forall i : p_i = \frac{1}{6}$ which has 0 d.o.f. is the best description of the die anyway.

The proposal favours many-parameter models; overfitting is avoided by other means. SED HANC MARGINIS NON CAPERET. . .

Conclusion

- We're examining a sample space with a prior
- Randomness comes from the Universe (when we make the experiment)
- (Potential) machine learning applications:
  - Parameter steering for SAXS/SANS modelling (Tanner)
    - No "simple" estimator gives the parameters $x \in X$
  - Model selection (Tanner): describing the prior in terms of the model, towards a learning based model
  - Hypothesis whittling: for $h \in H$ construct experiment that tests h, tests $\neg h$, or ignores h.
  - From Hypothesis Whittling to explainable AI
    - Inference of importance (from posterior) of hypotheses
  - Hypotheses and explainable AI
  - Testing *all* the data instead of the classical sample can test more hypotheses – better coverage.
    - Assuming of course that the tests work
- This talk has taken us from DLS to the philosophy of science to machine learning to some open questions and possible future directions. . .

Future Directions

Any, all, or none of these could be topics for a future talk:

- Information and Ω (the Universe) and entropy; KL divergence; hypotheses testing and subhypotheses testing (refinement).
- Hypotheses and model selections. Comparing Classical with Bayesian model selection.
- "Classical" loss functions (least squares, max likelihood) vs machine learning loss functions.
- Randomness for science [XKCD] . . . (vs. randomness for cryptography [NIST-IR6483])

References

- [3SAT] https://en.wikipedia.org/wiki/Boolean_satisfiability_problem
- [NIST-AI] R. Kuhn, R. Kacker: *An Application of Combinatorial Methods for Explainability in Artificial Intelligence and Machine Learning*, (draft) https://csrc.nist.gov/publications/detail/white-paper/2019/05/22/combinatorial-methods-for-explainability-in-ai-and-ml/draft
- [NIST-IR6483] J. Soto, L. Bassham: *Randomness Testing of the AES Finalist Candidates*, NIST (Apr 2000)
- [XKCD] https://xkcd.com/882/

Acks
Author funded by ATI grant. Typeset with amsLaTeX/Beamer/XYpic.