Voyages into Machine Learning

Jesse Thaler

based on: 1708.02949 with Eric Metodiev & Ben Nachman

1712.07124
1810.05165 with Eric Metodiev & Patrick Komiske

1802.00008
1809.01140 with Eric Metodiev + Patrick Komiske

I was initially trained as a BSM model builder, but over time I've become interested in methods to let the data "speak for itself." This includes model independent searches as well as a suite of generally-applicable jet substructure techniques.

Ultimately, our field is moving towards "anomaly detection", where we search for deviations from the Standard Model without specifying what BSM physics we are looking for. I don't know how to make that well defined, though, but I'm trying. (and why I'm on this boat)

A big change in my research philosophy came with my grad students: Eric Metodiev & Patrick Komiske.

They have more of a CS / data science background, and they really pushed me to think about how to define problems explicitly.

ML ≠ black box

ML = well-specified optimization problem.

But question is: what to optimize?!
(and subject to what constraints?)

More generally, ML yields functions defined by data.

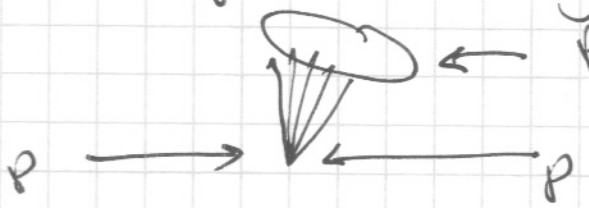
Supervised learning: functions defined by labelled training data.

Unsupervised learning: functions defined by data at hand.

Many problems can be approached with this philosophy but I'll focus on simplest: binary classification with case study of quark vs. gluon discrimination.

q vs. g : who cares?

First, quarks / gluons create jets in our detector



Showering, fragmentation, hadronization, detector effects.

q: $C_F = 4/3$ g: $C_A = 3$ ← ~twice as much radiation

If you can discriminate q vs. g (with some fidelity), you can improve, e.g. SUSY searches @ colliders

Signal: $pp \to \tilde{g}\tilde{g} \to q\tilde{q} \to q\tilde{\chi}$ ⇒ 4 quark jets

Background: QCD multijet ⇒ more gluon jets.

What is a quark or gluon jet?

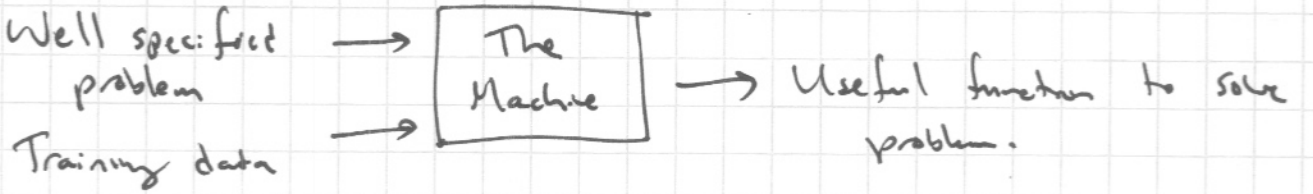
color triplet vs. color octet, but

hadrons are color neutral (hmm...)

How do you optimally distinguish quark from gluon? (assuming you could even define it...)

Remarkably, same answer: ML on mixed samples of quarks & gluons. (with many caveats)

Cartoon of Machine Learning



For quark/gluon: we want $h(\text{jet}) = \begin{cases} 1 & \text{for quark} \\ 0 & \text{for gluon} \end{cases}$

We have unlabeled jet samples.

And we have strong theory knowledge / biases.

First, best you can ever do is

$$h_{opt}(\text{jet}) = \frac{p(\text{jet} \mid q)}{p(\text{jet} \mid q) + p(\text{jet} \mid g)}$$

(Neyman-Pearson Lemma) Basically if there is a chance to get same configuration from both quark/gluon, you have irreducible overlap and can only guess. (with some chance of being right)

But with enough labelled training data, you can learn $h(\text{jet}) \approx h_{opt}(\text{jet})$ with gradient descent

$$\text{loss} = \left\langle \left(h(\text{jet}) - 1\right)^2 \right\rangle_{\text{quark}} + \left\langle \left(h(\text{jet}) - 0\right)^2 \right\rangle_{\text{gluon}}$$

↑ thing I'm trying to minimize.

↑ thing I'm trying to determine

average over samples of pure jets.

A few lines of algebra to show minimum is $h_{opt}(\text{jet})$.

Modern ML = fancy functional choices for $h(\text{jet})$ that are easy to minimize.

What is $h(\text{jet})$?

"jet" = set of 4-vectors, with flavor labels

↑ unordered, variable length list.

↑ Quantum Mechanics!

A very powerful parametrization of $h(\text{jet})$ that respects permutation symmetry:

$$h(\text{jet}) : h(l_1, l_2, \ldots l_L)$$

↑ ↑ ↑ ↑

"latent variables"

$$l_i = \sum_{\text{particles}} f(p_i)$$

↑ ↑ per particle

sum preserves symmetry

Really simple! Possibly complete if L large enough.

Easy to plot f and h, can gain intuition for physics.

But need laptop for that. (show psychedelic images?)

Big question: how do I find $h_{s/b}(\text{jet})$ if I don't have pure quark / gluon samples?!

For me, this is when I began to appreciate that ML is not really about the machine. It is about the physics.

Have mixed sample:

$$p(\text{jet}) = f_q \, p_q(\text{jet}) + f_g \, p_g(\text{jet})$$

Assume sample independence.

Assume you have access to samples with different quark/gluon fractions.

Sample A: $f_q = 1 - f_g = f_A$

Sample B: $f_q = 1 - f_g = f_B$

Optimally A vs. B discriminant:

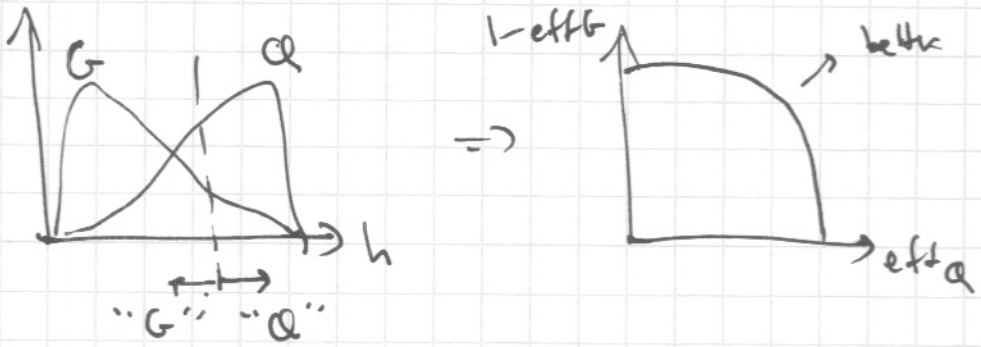
$$\frac{f_A \, p_q(\text{jet}) + (-f_A) \, p_g(\text{jet})}{(f_A + f_B) \, p_q(\text{jet}) + (2 - f_A - f_B) \, p_{\text{gluon}}(\text{jet})}$$

If you stare at this long enough....

h_{opt}^{mixed} is a function only of h_{opt}^{pure} ... and

$$\frac{\partial h_{opt}^{mixed}}{\partial h_{opt}^{pure}} = \frac{f_A - f_B}{(\text{something})^2}$$

So h_{opt}^{mixed} is monotonically related to optimal classifier!
(i.e. decision boundaries are the same)



Same ROC curve with monotonic scaling.

So train on mixed samples => optimal decision boundary

"CWoLa" = classification without labels.

We don't know purity of sample, though...
(i.e. what decision boundaries correspond to what purity)
But if we knew p_A and p_B we would be done!

Here is the magic: You can use classifier itself to define categories via "mutual irreducibility"

What is a "quark"? Hard to answer.

Is this particle-level jet a "quark"? Hard to answer.

Is there any example jet (even if very rare), where I can say definitely it is a quark? (anchor bin)

Example from text processing:

"energy" physics or climate?

"energy conservation" physics or climate?

"Noether's Theorem" / "Kyoto Protocol" definitely → physics / climate

Identify "most quark-like" region

But we already have good discriminant! Cut really hard, and $h_{opt}^{jet} \to 1$ and $h_{opt}^{jet} \to 0$ would be pure categories.

For h_{opt}^{mixed} $\longrightarrow$ $\left[\dfrac{1-f_A}{2\cdot f_A - f_B} \;,\; \dfrac{f_A}{f_A+f_B} \right]$

smallest largest

So basically endpoints of h_{opt}^{mixed} can be solved to find fractions.. (two equations, two unknowns)

Too good to be true? Yes, no stats at endpoints, but there are tricks to get around that.

What if you have three samples? Then you will find three "topics" (mutually irreducible distributions where each category is pure in some "anchor bin".

If only two underlying categories, you'll find degenerate distributions (which you can quantify).

Bottom line: Given mixed samples, can use ML to define underlying categories.

Data-driven bootstrap. (aka. magic)

Looking ahead: Anomaly Detection.

Want to find deviations from Standard Model.

If you have trustable MC, you can build

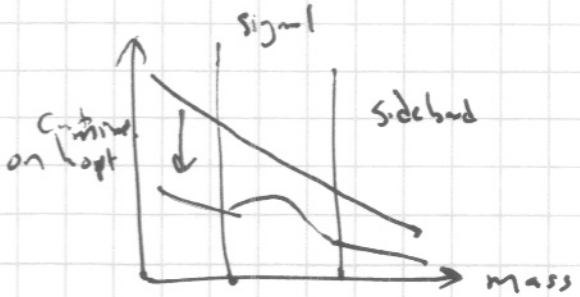
$$h_{opt}(\text{event}) : \frac{p(\text{event}\,|\,\text{data})}{p(\text{event}\,|\,\text{data}) + p(\text{event}\,|\,\text{MC})}$$

and any deviations from 1/2 would be evidence for new physics

But we don't have a trustable MC always!

First attempt at solution: CWoLa Hunting.

Collins, Howe, Nachman 1805.02664



Build optimal signal/sideband discriminant, cut hard to maybe expose bump

This assumed we knew that new physics would show up as bump. What if we don't know which feature separates signal from background?

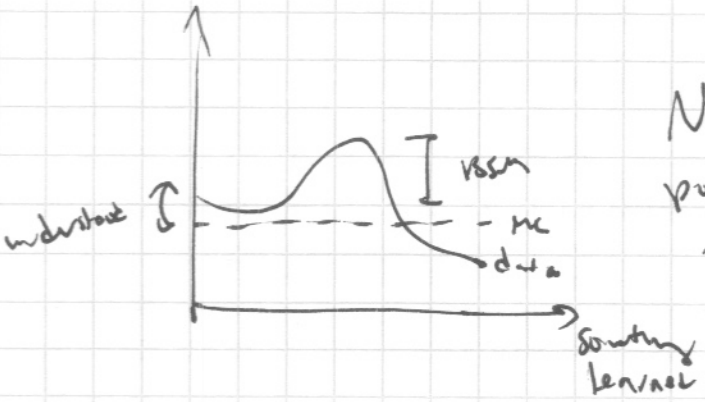
(Hence 8 physicists on a boat)

The problem: I have $p(\text{event}|MC)$ but it isn't perfect.

I think new physics will show up in some "localized" way as a generalized bump.

How do I find localized features in $p(\text{event}|\text{data})$ that differs from $p(\text{event}|MC)$?

Localized, since less likely to be misunderstood MC effect.



Not clear that this is possible, but now that I know topological modeling works, inclined to try.