

From traditional statistical modelling to machine learning algorithms:

a perspective from the STRATOS TG9 high-dimensional data viewpoint

Riccardo De Bin

Department of Mathematics – University of Oslo

Co-authors: Federico Ambroggi, Lara Lusa, Jörg Rahnenführer, Willi Sauerbrei

Overview: this talk's perspective



*"[Statistics and Machine Learning] are like two pairs of old men sitting in a park playing two different board games. Both games use the **same type of board** and the **same set of pieces**, but each **plays by different rules** and **has a different goal** because the games are fundamentally different. Each pair looks at the other's board with bemusement and thinks they're not very good at the game."*

(Fawcett & Hardin, 2017)

Outline of the lecture

- Statistics and Machine Learning
 - Same type of board
 - Same set of pieces
 - Play by different rules
- Biomedical perspective
- STRATOS TG9
- Final remarks

Same type of board: $f : \mathcal{X} \rightarrow \mathcal{Y}$

Let:

- $X \in \mathcal{X}$ be the inputs;
- $Y \in \mathcal{Y}$ be the outcome.

In the supervised learning framework, statistics and machine learning try to find the best $f : \mathcal{X} \rightarrow \mathcal{Y}$ to relate the inputs X to the outcome Y ,

$$X \xrightarrow{f(X)} Y,$$

such that we can estimate an $\hat{Y} = \hat{f}(X)$ for a given value of X .

How to define this “best” is a matter of rules and goals . . .

Same type of board: example

Example of (extremely simple) rule:

$$f(x) = 2 + 1.5x - 0.4x^2.$$

This rule can be found both by implementing:

- a **supposedly** “statistical approach”, like a fractional polynomial,

$$\text{FP}(m) : f(x) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} + \dots + \beta_m x^{p_m},$$

with $m = 2$ and $(p_1, p_2) = (1, 2)$; so $\beta_0 = 2$, $\beta_1 = 1.5$, $\beta_2 = -0.4$.

- a **supposedly** “machine learning approach”, like a neural network,

$$\text{NN}(k) : f(x) = b^{(2)} + \sum_{k=1}^K w_k^{(2)} \phi(w_k^{(1)} x + b_k^{(1)}),$$

with $\phi(x) = x^2$ and $k = 1$; so that $b^{(1)} = -1.875$, $w^{(1)} = 1$,
 $b^{(2)} = 3.40625$, $w^{(2)} = 0.4$.

Same type of board and same set of pieces

Not only do statistics and machine learning operate on the same board (space of functions $f(X)$), but their models/rules $\hat{f}(X)$ are often just **optimisers**

$$\hat{f}(X) = \arg \min_f E [L(Y, f(X))],$$

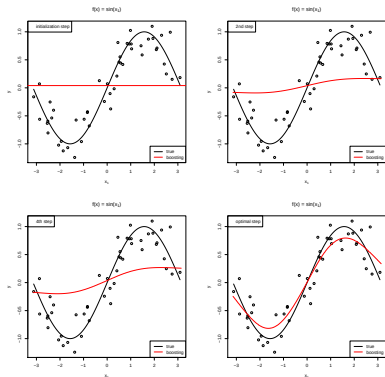
where $L(Y, f(X))$ is a **loss function** that penalises errors (sum of squares, L_2 loss, negative log-likelihood, ...):

- not only the final product (estimate) $\hat{f}(x)$ may be the same;
- the optimiser itself may be the same...

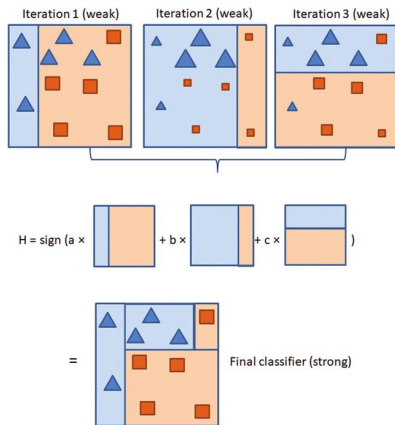
Same set of pieces: example

Consider a classification problem ($Y \in \{-1, 1\}$) and a boosting algorithm:

Forward Stagewise Additive Model



AdaBoost



(Tang et al., 2021, Figure 2)

Same set of pieces: example

Despite being conceptually different, the two algorithms turn out to be **equivalent** when minimising the exponential loss:

- the final classifier is based on one-half of the log-odds of $P(Y = 1|x)$ (Hastie et al., 2009, Section 10.5).



Starting from here, a large number of boosting algorithms have been developed (e.g., GBM, mboost, GAMBoost, XGBoost, LightGBM, APLR, CatBoost, gamboostLSS, SMARTboost, ...):

- setting a line between statistics and machine learning is **impossible and ultimately pointless**;
- the focus should be on **how** they are used ...

Play by different rules: statistics and machine learning

What do we mean by “play with different rules” here?¹

Statistics

- “The goal of modeling is approximating and then understanding the **data-generating process**”
- “The models provide the **mathematical framework**”
- “Each choice made in the analysis must be **defensible**”
- “The **analysis** is the final product”

Machine Learning

- “The predominant task is **predictive** modeling”
- “The model is really only instrumental to its **performance**”
- “The proof of the model is in the **test set**”
- “**Free** from worrying about model assumptions or diagnostics”

¹all these quotes are from Fawcett & Hardin (2017)

Play by different rules: statistics and machine learning

Strong connections with Breiman (2001):

- Should our model be a simplified representation of the **data-generating process** useful to answer a research question?
- Or should we abandon the attempt to understand the world's complexity and **focus on the result** of the complex process?

A special issue re-evaluating Breiman's "Two Cultures" framework after 20 years has been published in *Observational Studies*:

- many interesting contributions and reflections;
- "tomorrow's successful data analyst will be well-versed in the classical data modeling toolkit, **as well as** in algorithmic modeling approaches" (Neufeld & Witten, 2021)

Play by different rules: example

Consider using an ensemble method:

Statistics

- multi-model inference (e.g., BMA, FMA, ...);
- **reliable confidence/credible intervals**, based on asymptotic theory;
- strong assumptions, need for diagnostic tools;
- prediction is possible, not necessarily the main goal.

Machine Learning^a

- fit multiple CARTs to data reshuffles (e.g., random forests);
- the theory behind their functioning is well understood;
- no assumptions on when they can be used;
- produce unintelligible rules (black box);
- **amazingly successful at prediction tasks.**

^afrom Fawcett & Hardin (2017)

Biomedical perspective: we usually need both

Task	Inferential component	Predictive component
Drug development	Estimate treatment effects and uncertainty	Identify patients likely to benefit
Biomarker research	Understand biological association/confounding	Build a risk or diagnostic score
Precision medicine	Estimate heterogeneous treatment effects	Forecast response for an individual
Public health	Explain determinants of disease	Forecast burden and allocate resources

Biomedical perspective: example

Suppose we observe, for each patient:

- outcome Y : survival, treatment response, toxicity, ...
- predictors X : age, biomarkers, imaging, gene expression, ...
- treatment/exposure A : treatment received or risk factor,

and we want to use clinical and genomic variables to **improve care** for patients with cancer.

We can use different tools to answer **different** kinds of questions:

Statistical tools

- Does treatment A improve survival?
- What is the effect size, its uncertainty, and its interpretation?

Machine Learning tools

- Which patient is likely to respond?
- How accurately can we predict the outcome for a new patient?

STRATOS TG9: handling machine learning

In STRATOS, the issues related to machine learning are handled by the
TG9 High-dimensional data:

- seems natural, as in high-dimensional frameworks the differences are even **less** clear (think, for example, of checking assumptions);
- many **common issues** (some even to other frameworks):
 - ▶ clear definition of target population and intended use;
 - ▶ difficulties in defining a sufficient sample size;
 - ▶ protection against overfitting;
 - ▶ clinically meaningful performance metrics;
 - ▶ reproducibility and reporting;
- currently working on both **systematic reviews** and **guidance papers**.

Conclusions

Final remarks:

- “Both Statistics and Machine Learning create *models from data*, but for *different purposes*” (Fawcett & Hardin, 2017);
- there is *no statistics/machine learning dichotomy*: the appropriate tool depends on the scientific question and the problem at hand;
- both statistics and machine learning models require *clear intended use, generalisability, reproducibility, transparency, and clinical utility*;
- the focus should not be on “statistics versus machine learning” but on “how to build *trustworthy and clinically useful models*”.

References I

- BREIMAN, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* **16**, 199–231. Publisher: Institute of Mathematical Statistics.
- FAWCETT, T. & HARDIN, D. (2017). Machine Learning vs. Statistics.
<https://www.svds.com/machine-learning-vs-statistics/>.
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition)*. Springer, New York.
- NEUFELD, A. & WITTEN, D. (2021). Discussion of Breiman's “two cultures””: From two cultures to one. *Observational Studies* **7**, 171–174.
- TANG, J., HENDERSON, A. & GARDNER, P. (2021). Exploring adaboost and random forests machine learning approaches for infrared pathology on unbalanced data sets. *Analyst* **146**, 5880–5891.

THANK YOU!

"I don't know what Machine Learning will look like in ten years, but whatever it is I'm sure Statisticians will be whining that they did it earlier and better" (Fawcett & Hardin, 2017)