LEARNING NEW PHYSICS FROM A MACHINE

Raffaele Tito D’Agnolo - SLAC
CIPANP 2018

Raffaele Tito D’Agnolo and Andrea Wulzer
To Appear (next week)
\( \chi^2 = 47 \quad N_{\text{bins}} = 50 \quad p - \text{value} < 1\sigma \)
THE PROBLEM

\[ t_{id}(\mathcal{D}) = 2 \log \left( \frac{e^{-N(\text{NP})}}{e^{-N(\text{R})}} \prod_{x \in \mathcal{D}} \frac{n(x|\text{NP})}{n(x|\text{R})} \right) \]
WHAT IS A NEURAL NETWORK?

SET OF FUNCTIONS

+ FITTING ALGORITHM
WHAT IS A NEURAL NETWORK?

SET OF FUNCTIONS

\( f^{(1)}_{w_1} \left( f^{(2)}_{w_2} \left( f^{(3)}_{w_3} (\ldots) \right) \right) \)
1. LINEAR TRANSFORMATION  \( z = \vec{w} \cdot \vec{x} + b \)  
FREE PARAMETERS

2. NON-LINEAR TRANSFORMATION  \( \sigma(z) \)  
FIXED
2. NON-LINEAR TRANSFORMATION

\[ \sigma(z) = \begin{cases} 
\tanh(z) & \text{tanh(z)} \\
\text{ReLU} & \frac{1}{1+e^{-z}} \\
& \ldots
\end{cases} \]
\[
\sigma \left( \sum_{k=1}^{3} w_{jk} \sigma_k \left( \sum_{i=1}^{3} w_{ki} \sigma_i \left( \sum_{l=1}^{d} w_{il} x_l + b_i \right) + b_k \right) + b_j \right)
\]
UNIVERSAL APPROXIMANTS

Increasing $w$

$$w, w_1, w_2$$

Height

$w_1 \sigma_1 + w_2 \sigma_2 + b'$

$w_1 = -w_2$

Width $b$

$$\sigma(wx + b)$$
Fitting Algorithm

Loss Function

\[ L = \frac{1}{N_c} \sum_{i=1}^{N_c} [1 - f_{NN}(\tilde{x}_i, w, b)]^2 + \frac{1}{N_d} \sum_{j=1}^{N_d} [f_{NN}(\tilde{x}_j, w, b)]^2 \]

Training

\[ w_{t+1} \rightarrow w_t - \epsilon \partial_w \hat{L} \]

\( \hat{L} \) SUBSET OF THE SAMPLE

\( \epsilon \) LEARNING RATE
LEARNING NEW PHYSICS
A SIMPLE STRATEGY

1. LEARN THE DATA DISTRIBUTION

\[ n(x|\hat{w}) \approx n(x|T) \]
A SIMPLE STRATEGY

1. LEARN THE DATA DISTRIBUTION  
   \( n(x|\hat{w}) \approx n(x|T) \)

2. CHECK IF IT IS DIFFERENT FROM THE REFERENCE ONE

   \[
   t(D) = 2 \log \left[ \frac{e^{-N(\hat{w})}}{e^{-N(R)}} \prod_{x \in D} \frac{n(x|\hat{w})}{n(x|R)} \right] \\
   p_{\text{obs}} = \int_{t_{\text{obs}}}^{\infty} dt P(t|R)
   \]

STANDARD LIKELIHOOD RATIO
NEYMAN-PERSON TEST STATISTIC

REFERENCE DISTRIBUTED TOYS
\[ t(D) = 2 \log \left[ \frac{e^{-N(\hat{w})}}{e^{-N(R)}} \prod_{x \in D} \frac{n(x|\hat{w})}{n(x|R)} \right] \]

\[ = -2 \min_{\{w\}} \left[ \frac{N(R)}{N_\mathcal{R}} \sum_{x \in \mathcal{R}} (e^{f(x;w)} - 1) - \sum_{x \in D} f(x;w) \right] \]

\[ \equiv -2 \min_{\{w\}} L[f(\cdot, w)] \]
THE LOSS FUNCTION

\[ t(D) = 2 \log \left( \frac{e^{-N(\hat{w})}}{e^{-N(R)}} \prod_{x \in D} \frac{n(x|\hat{w})}{n(x|R)} \right) \]

\[ = -2 \min_{\{w\}} \left[ \frac{N(R)}{N(R)} \sum_{x \in R} (e^{f(x;w)} - 1) - \sum_{x \in D} f(x;w) \right] \]

THE NETWORK IS DOING A MAXIMUM LIKELIHOOD FIT TO THE DATA AND COMPUTING THE "OPTIMAL" TEST STATISTIC AT THE SAME TIME
SENSITIVE TO NEW PHYSICS

- $P(t|R)$
- $P(t|NP_1)$

4 Neurons
Peak in the Tail
No cut

$\chi^2_{13}$

Peaks in the Tail, 4 Neurons, No cut

Median NN

Median Ideal
INSENSITIVE TO CUTS

Reference
No cut
x>0.3
x>0.5
x>0.65

NP1
No cut
x>0.3
x>0.5
x>0.65
MODEL-INDEPENDENT

- Diagrams showing distributions of $P(t|R)$ for 4 Neurons with excess in the tail and peak in the bulk, respectively.
- Each diagram includes fiducial and non-fiducial (NP) cuts.
- The $Z_{id}$ and $Z_{NN}$ axes are plotted for the ideal and median case.
CONCLUSION AND OUTLOOK

• TODAY IN FUNDAMENTAL PHYSICS WE HAVE LARGE, MULTIVARIATE, SM-LIKE
  DATASETS AND STRONG REASONS TO BELIEVE THAT THEY SHOULD NOT BE SM-
  LIKE

• OUR BEST GUESSES FOR NEW PHYSICS ARE NOT BEING DETECTED AND
  ANYTHING THAT HELPS US TO SEARCH WITHOUT ANY BIAS CAN BE USEFUL

• NEURAL NETWORKS ARE WIDELY USED TO APPROXIMATE PROBABILITY
  DISTRIBUTIONS AND ARE IDEAL CANDIDATES FOR THIS TYPE OF PROBLEM

• TODAY I HAVE DESCRIBED AN APPLICATION OF NEURAL NETWORKS,
  FOUNDED ON SOLID STATISTICAL PRINCIPLES, WHICH GOES IN THIS DIRECTION

  • ITS VIRTUES (SENSITIVITY TO NP, MODEL-INDEPENDENCE, INSENSITIVITY TO
    CUTS) HAVE BEEN TESTED ON SIMPLE 1D AND 2D EXAMPLES

  • MORE WORK IS NEEDED IN THE 2D AND HIGHER-DIMENSIONAL CASE
BACKUP
TWO DIMENSIONS

NP: $x \sim \text{EXPONENTIAL} + \text{PEAK}$

$y \sim \text{UNIFORM}$

R: $x \sim \text{EXPONENTIAL}$

$y \sim \text{UNIFORM}$

RECOVERS COMPARABLE SENSITIVITY TO 1D FOR $x > 0.3$ OR DOUBLING THE EVENTS