Machine Learning for High Energy Physics

Andrea De Simone

andrea.desimone@sissa.it
The APP group at SISSA

interdisciplinary group working at interface of particle physics, astrophysics and cosmology

Address fundamental issues about our Universe: origin and evolution, nature of gravity, properties of dark matter and dark energy.

PIs in the group:
A. De Simone, S. Liberati, P. Ullio, M. Viel

PIs affiliated from other SISSA groups:
C. Baccigalupi, R. Percacci, S. Petcov, A. Romanino, P. Salucci

PIs affiliated from other institutions
P. Creminelli, E. Sefusatti
Speech recognition
Recommender systems
Creative Paintings

Which of these images were created by a machine?
Generating Faces

These people do not exist!

CelebA-HQ:  https://youtu.be/XOxxPcy5Gr4

[Karras et al - 1710.10196]
Autonomous driving
Games

Google’s DeepMind plays Breakout

AlphaGo beats world champion Go 4-1
2010-05-06, 2:45pm
2010-05-06, 2:45pm

Flash Crash!

Lost ~9% in 36 minutes!

Dow-Jones Industrial Average
> Outline

1. Machine Learning in Science

2. Open problems in High-Energy Physics (HEP)

3. Statistical test of dataset compatibility

4. Applications to HEP
1. Machine Learning in Science

2. Open problems in High-Energy Physics (HEP)

3. Statistical test of dataset compatibility

4. Applications to HEP
“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

[Mitchell - 1997]
Why is Machine Learning so cool?

- a guide through big data (data mining)
- many diverse applications
  (from engineering to commerce to science)
- can help making our life better/easier
- …
- can help scientists to do science better and faster
> Why now?

- **More powerful machines**
  
  both speed and storage

- **More data**

  almost everything is recorded!

- **Easier access**

  internet revolution, easier to share big data
Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
 Supervised Learning

labelled data
\( \{(x_i, y_i)\}_{i=1}^{N} \)

features
labels

Logistic Regression
Neural Networks
Decision Trees
Nearest Neighbors
...

Polynomial Regression
Neural Networks
Support Vector Machines
Nearest Neighbors
...

machine “learns” the model

\[ f(x) = y \]
Unsupervised Learning

Cluster Analysis
Dimensionality Reduction
Anomaly Detection
...

unlabelled data
\{x_i\}_{i=1}^N

features

machine "learns" patterns, structures, representations, etc. of the data
deep learning achieved state-of-the-art results.

- Mammography
- Brain lesions
- Airways
- Diabetic retinopathy
- Prostate
- Skin lesions
- Lung nodule
- Bone suppression

[Litjens et al. - Medical Image Analysis 2017]
ML in Medical Science

Dermatologist-level classification of skin cancer with deep neural networks

sensitivity = \frac{True	ext{ Positives}}{Positives}

specificity = \frac{True	ext{ Negatives}}{Negatives}

[Esteva et al. - Nature 2017]
Neural Networks for the prediction of organic chemistry reactions

predict probability of 17 different reaction types

(≈85% accuracy on test set)
Detect extreme weather using deep learning

classification accuracy

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Train</th>
<th>Test</th>
<th>Train time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical Cyclone</td>
<td>99%</td>
<td>99%</td>
<td>≈ 30 min</td>
</tr>
<tr>
<td>Atmospheric River</td>
<td>90.5%</td>
<td>90%</td>
<td>6-7 hour</td>
</tr>
<tr>
<td>Weather Front</td>
<td>88.7%</td>
<td>89.4%</td>
<td>≈ 30 min</td>
</tr>
</tbody>
</table>

[Liu et al - 1605.01156]
Machine learning phases of matter

square-lattice Ising model:
\[ T_c/J = 2.266 \pm 0.002 \]
(exact: 2.269..)

triangular-lattice Ising model:
\[ T_c/J = 3.65 \pm 0.01 \]
(exact: 3.64095..)

[Carrasquilla, Melko - Nature 2017]
Figure 5: (top) ROC and (bottom) SIC curves of the FLD and the deep convolutional network trained on (left) 200 GeV and (right) 1000 GeV Pythia jet images with and without color compared to baseline jet observables and a BDT of the five jet observables.

- Efficiency at 50% quark jet classification efficiency for each of the jet variables and the CNN are listed in Table 1.
- To combine the jet variables into more sophisticated discriminants, a boosted decision tree (BDT) is implemented with scikit-learn. The convolutional network outperforms the traditional variables and matches or exceeds the performance of the BDT of all of the jet variables. The performance of the networks trained on images with and without color is shown in Figure 6.

5.1 Colored Jet Images

The benchmarks in the previous section were compared to the jet images with and without color, where the three color channels correspond to separating out the charge and multiplicity information as described in Section 3.3. Figure 6 shows the SIC curves of the neural network performances with and without color on Pythia jet images. For the 100 GeV and 200 GeV images, only small changes in the network performance were observed by adding in color of this form. For the 500 GeV and 1000 GeV jet images, performance increases were consistently observed.

Figure 2: An illustration of the deep convolutional neural network architecture. The first layer is the input jet image, followed by three convolutional layers, a dense layer and an output layer. The max-pooling layers performed a 2 \times 2 down-sampling with a stride length of 2. The dense layer consisted of 128 units.

- All neural network architecture training was performed with the Python deep learning libraries Keras and Theano on NVidia Tesla K40 and K80 GPUs using the NVidia CUDA platform. The data consisted of the 100k jet images per \( p_T \)-bin, partitioned into 90k training images and 10k test images. An additional 10% of the training images are randomly withheld as validation data during training of the model for the purposes of hyperparameter optimization. He-uniform initialization was used to initialize the model weights. The network was trained using the Adam algorithm using categorical cross-entropy as a loss.

[Komiske, Metodiev, Schwartz - 1612.01551]
> Outline

1. Machine Learning in Science

2. Open problems in High-Energy Physics (HEP)

3. Statistical test of dataset compatibility

4. Applications to HEP
The Standard Model

One of the most accurately verified predictions in history of physics

anomalous magnetic moment of the electron

\[ a_{\text{exp}} = 0.001\, 159\, 652\, 180\, 91(26) \]

\[ a_{\text{th}} = 0.001\, 159\, 652\, 181\, 643(764) \]

Excellent description of sub-nuclear phenomena

One of the most accurately verified predictions in history of physics

more than 10 significant figures!
Is there anything left to discover?
Need to go beyond...

- Neutrino Masses
- Matter/Anti-matter Asymmetry
- Dark Matter
- Strong-CP problem
- Dark Energy
- Hierarchy problem

Standard Model
Large Hadron Collider (LHC)

The most complex (and expensive) experiment ever built!

Cost ~ 4 GigaEur

counter-rotating proton beams in 27km circumference ring

center-of-mass energy: 13 TeV

d Detectors at 4 collision points:
- ATLAS
- CMS
- LHCb
- ALICE
> ATLAS experiment

~ 2100 physicists
37 countries
167 universities/labs
Signal vs Background

The figure illustrates the concept of signal and background events in data analysis. The left side shows distributions of two variables: Mass (GeV/c^2) for the first and Alternate Variable for the second. Each distribution is divided into cut events and keep events, with signal events highlighted in red and background events in blue. The figure aims to simplify the identification of signal from background by using visual aids such as magnifying glasses and a magnet.
Standard Analysis Pipeline

Bkg & Signal Simulation -> Events Selection -> Variables Selection -> Histograms

Expected counts -> Statistical Significance ($\chi^2$, likelihood ratio, …) -> Observed counts

Data
Standard Analysis Pipeline

Model 1
- Data
- Standard Analysis
- compatible?
  - yes: discovery!
  - no: constraints on params

Model 2
- Data
- Standard Analysis
- compatible?
  - yes: discovery!
  - no: constraints on params

... 

Model N
- Data
- Standard Analysis
- compatible?
  - yes: discovery!
  - no: constraints on params
New Physics?

Searches for New Physics Beyond the Standard Model have been negative so far...

MAYBE:

1. New Physics is not accessible by LHC
   new particles are too light/heavy
   or interacting too weakly

2. We have not explored all the possibilities
   new physics may be buried under large bkg
   or hiding behind unusual signatures
“Don’t want to miss a thing” (in data)

closer look at currently available data
get ready for upcoming data from next Run of LHC

Model-independent search

searches for specific models may be insensitive
to unexpected / unknown / anomalous processes
1. Machine Learning in Science

2. Open problems in High-Energy Physics (HEP)

3. Statistical test of dataset compatibility

4. Applications to HEP
Want a statistical test for NP which is:

1. **model-independent:**
   no assumption about underlying physical model to interpret data
   more general

2. **non-parametric:**
   compare two samples as a whole (not just their means, etc.)
   fewer assumptions, no max likelihood estim.

3. **un-binned:**
   high-dim feature space partitioned without rectangular bins
   retain full multi-dim info of data
Two-sample Test

Two sets:

Trial: \( \mathcal{T} = \{x_1, \ldots, x_{N_T}\} \overset{iid}{\sim} p_T \)

Benchmark: \( \mathcal{B} = \{x'_1, \ldots, x'_{N_B}\} \overset{iid}{\sim} p_B \)

probability distributions \( p_B, p_T \) unknown

e.g.: simulated SM bkg real measured data

[a.k.a. “homogeneity test”]
> Two-sample Test

Two sets:

Trial: \[ \mathcal{T} = \{ x_1, \ldots, x_{N_T} \} \overset{iid}{\sim} p_T \]

Benchmark: \[ \mathcal{B} = \{ x'_1, \ldots, x'_{N_B} \} \overset{iid}{\sim} p_B \]

Probability distributions \( p_B, p_T \) unknown

Are \( B, T \) drawn from the same prob. distribution?

easy…
Two-sample Test

Two sets:

Trial:  \( \mathcal{T} = \{ x_1, \ldots, x_{N_T} \} \overset{iid}{\sim} p_T \)

Benchmark:  \( \mathcal{B} = \{ x'_1, \ldots, x'_{N_B} \} \overset{iid}{\sim} p_B \)

probability distributions \( p_B, p_T \) unknown

Are \( B, T \) drawn from the same prob. distribution?

… hard!
Why is it important?

- decide whether two datasets can be analyzed jointly
- find anomalous data points (outliers)
- detect changes in the underlying distributions
- detect events in streams of data (time-series data)
- check if data are compatible with expectations
- ...

> Two-sample Test
Two-sample Test

RECIPE:

1. Density Estimator
   → reconstruct PDFs from samples

2. Test Statistic (TS)
   → measure “distance” between PDFs

3. TS distribution
   → associate probabilities to TS
   under null hypothesis $H_0: p_B = p_T$

4. $p$ -value
   → accept/reject $H_0$
> 1. Density Estimator

Divide the space in squared bins?

✓ easy
✓ can use simple statistics (e.g. $\chi^2$)
✗ hard/slow/impossible in high-$D$

Need un-binned multivariate approach

Find PDFs estimators: $\hat{p}_B(x), \hat{p}_T(x)$
e.g. based on densities of points:

$$\hat{p}_{B,T}(x) = \frac{\rho_{B,T}(x)}{N_{B,T}}$$

Nearest Neighbors!

[Schilling - 1986][Henze - 1988]
[Wang et al. - 2005, 2006]
[Dasu et al. - 2006][Perez-Cruz - 2008]
[Sugiyama et al. - 2011][Kremer et al, 2015]
1. Density Estimator

- Fix integer $K$.
- Choose query point $x_j$ in $T$ and draw it in $B$. 
> 1. Density Estimator

- Fix integer $K$.
- Choose query point $x_j$ in $T$ and draw it in $B$.
- Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$. 
1. Density Estimator

- Fix integer $K$.
- Choose query point $x_j$ in $T$ and draw it in $B$.
- Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$.
- Find the distance $r_{j,T}$ of the $K^{th}$-NN of $x_j$ in $T$. 

$B$ 

$T$ 

$X_j$ 

$r_{j,B}$ 

$r_{j,T}$ 

$X_j$
> 1. Density Estimator

- Fix integer $K$.
- Choose query point $x_j$ in $T$ and draw it in $B$.
- Find the distance $r_{j,B}$ of the $K^{th}$-NN of $x_j$ in $B$.
- Find the distance $r_{j,T}$ of the $K^{th}$-NN of $x_j$ in $T$.
- Estimate PDFs:

\[
\hat{p}_B(x_j) = \frac{K}{N_B} \frac{1}{\omega_D r_{j,B}^D}
\]
\[
\frac{1}{N_T - 1} \omega_D r_{j,T}^D
\]

\[
\hat{p}_T(x_j) = \frac{K}{N_T - 1} \frac{1}{\omega_D r_{j,T}^D}
\]
> 2. Test Statistic

- Measure of the “distance” between 2 PDFs

- Define **Test Statistic**:
  (detect under-/over-densities)

\[
TS(\mathcal{T}) \equiv \frac{1}{N_T} \sum_{j=1}^{N_T} \log \frac{\hat{p}_T(x_j)}{\hat{p}_B(x_j)}
\]

- Related to Kullback-Leibler divergence as:
  \[TS(\mathcal{T}) = \hat{D}_{KL}(\hat{p}_T||\hat{p}_B)\]

\[
D_{KL}(p||q) \equiv \int_{\mathbb{R}^D} p(x) \log \frac{p(x)}{q(x)} \, dx
\]

- From NN-estimated PDFs:
  \[TS(\mathcal{T}) = \frac{D}{N_T} \sum_{j=1}^{N_T} \log \frac{r_{j,B}}{r_{j,T}} + \log \frac{N_B}{N_T - 1}\]

- **Theorem**: this estimator converges to \(D_{KL}(\rho_B||\rho_T)\), in large sample limit

[Wang et al. - 2005, 2006]
### 3. Test Statistic Distribution

**How is TS distributed?**  **Permutation test!**

Assume $p_B = p_T$. Union set: $U = T \cup B$

1. Random reshuffle $U$
2. Compute the test statistic $TS_n$ on: $(\widetilde{B}, \widetilde{T})$

Repeat many times.

**Distribution of TS under $H_0$:** $f(TS|H_0) \leftarrow \{TS_n\}$

[asymptotically normal with zero mean]
> 4. $p$-value

- $\hat{\mu}, \hat{\sigma}$: mean, variance of TS distribution $f(TS|H_0)$

- Standardize the TS: $TS \rightarrow TS' \equiv \frac{TS - \hat{\mu}}{\hat{\sigma}}$

- TS’ distributed according to $f'(TS'|H_0) = \hat{\sigma} f(\hat{\mu} + \hat{\sigma} TS'|H_0)$

- Two-sided $p$-value:
  \[ p = 2 \int_{|TS'_{obs}|}^{+\infty} f'(TS'|H_0) dTS' \]

- Equivalent significance: $Z \equiv \Phi^{-1}(1 - p/2)$
Gaussian Example

\[ p_B = \mathcal{N}(\mu_B, \Sigma_B) \quad \quad p_T = \mathcal{N}(\mu_T, \Sigma_T) \]

\[ \mu_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad \mu_T = \begin{pmatrix} 1.2 \\ 1.2 \end{pmatrix} \]

\[ \Sigma_B = \Sigma_T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]

exact KL divergence

\[ \mu_B = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \quad \mu_T = \begin{pmatrix} 1.15 \\ 1.15 \end{pmatrix} \]

more data, more power

\[ K = 5, N_{\text{perm}} = 1000 \]

\[ Z = 5\sigma \]
Where are the discrepancies?

**Bonus: Characterize regions with significant discrepancies**

1. "Score" field over $T$: 
   \[
   Z(x_j) \equiv \frac{u(x_j) - \bar{u}}{\sigma_u}
   \]
   with: 
   \[
   u(x_j) \equiv \log \frac{r_{j,B}}{r_{j,T}}
   \]
   \[
   TS_{\text{obs}} = D \bar{u} + \text{const}
   \]

2. Identify points where $Z(x) > c$
   They contribute the most to large $TS_{\text{obs}}$
   → high-discrepancy (anomalous) regions

3. Apply a clustering algorithm to group them
> NN2ST: Summary

**INPUT:**

Trial sample: \( \mathcal{T} = \{ x_1, \ldots, x_{NT} \} \overset{iid}{\sim} \rho_T \)

Benchmark sample: \( \mathcal{B} = \{ x'_1, \ldots, x'_{NB} \} \overset{iid}{\sim} \rho_B \)

- \( K \): number of nearest neighbors
- \( N_{perm} \): number of permutations

**OUTPUT:**

\( \rho \)-value of the null hypothesis \( H_0: \rho_B = \rho_T \)

[check compatibility between 2 samples]
> **NN2ST: Summary**

K-NN density ratio estimation

**Test Statistic** $\text{TS}_{\text{obs}}$

<table>
<thead>
<tr>
<th>Benchmark sample</th>
<th>Trial sample</th>
</tr>
</thead>
</table>

$\text{TS}_{\text{obs}}$

TS distribution

$-|\text{TS}_{\text{obs}}|, \text{TS}, |\text{TS}_{\text{obs}}|$

$p$ value

Python code: [github.com/de-simone/NN2ST](github.com/de-simone/NN2ST)

Paper: [DS, Jacques - 1807.06038](1807.06038)
NN2ST: Summary

✓ general, model-independent
✓ fast, no optimization
✓ sensitive to unspecified signals
✓ useful when no variable can separate sig/bkg
✓ helps finding signal regions, optimal cuts, …

✘ need to run for each sample pair
✘ permutation test is bottleneck
✘ limited by sample accuracies
1. Machine Learning in Science

2. Open problems in High-Energy Physics (HEP)

3. Statistical test of dataset compatibility

4. Applications to HEP
Standard Analysis Pipeline

Bkg & Signal Simulation → Events Selection → Variables Selection → Histograms

Data

Expected counts → Statistical Significance ($\chi^2$, likelihood ratio, ...) → Observed counts
Our Method

Bkg Simulation (Benchmark)

Data (Trial)

No signal in data

Reject null hypothesis?

yes

hint of new physics!

select regions to explore

No signal in data

Reject null hypothesis?

no
**DM search @ LHC**

- “proof-of-principle” study
- bkg: \( Z \rightarrow \nu \bar{\nu} + (1, 2) j \)
  - sub-leading bkgs not included
- no full detector effects

**Benchmark:** \( \text{BKG}_1 \)

**Trials:** \( \text{BKG}_2 + \text{SIG} \)

\[ K = 5 \]

\[ N_{perm} = 3000 \]

**8 features:**
- n. of jets
- \( p_T, \eta \) of 2 leading jets
- \( E_T^{\text{miss}}, H_T \)
- \( \Delta \phi_{E_T^{\text{miss}}, j_1} \)

**DM + Z’ vector mediator**

\[ m_{DM} = 100 \text{ GeV} \]
\[ m_{Z'} = 1.2, 2, 3 \text{ TeV} \]
\[ g_{DM} = 1, g_q = 0.1 \]
\[ \sqrt{s} = 13 \text{ TeV} \]
> DM search @ LHC

B: $\text{BKG}_1$ (20k events)

T1: $\text{BKG}_2$ (20k events) + $\text{SIG}_1$ (2010 events)

T2: $\text{BKG}_2$ (20k events) + $\text{SIG}_2$ (375 events)

T3: $\text{BKG}_2$ (20k events) + $\text{SIG}_3$ (59 events)

<table>
<thead>
<tr>
<th>Sample</th>
<th>$M_{Z'}$</th>
<th>$\sigma_{\text{signal}}$</th>
<th>$Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.2 TeV</td>
<td>20.4 pb</td>
<td>$&gt;15 \sigma$</td>
</tr>
<tr>
<td>T2</td>
<td>2 TeV</td>
<td>3.8 pb</td>
<td>10 $\sigma$</td>
</tr>
<tr>
<td>T3</td>
<td>3 TeV</td>
<td>0.6 pb</td>
<td>0.13 $\sigma$</td>
</tr>
</tbody>
</table>

$N_{\text{sig}} = N_B \times \frac{\sigma_{\text{signal}}}{\sigma_{\text{bkg}}}$

- in real world: expect degradation of results (uncertainties)
- the method has value, it is worth exploring

Expt. Collab. at CERN interested in applying this test
> DM search @ LHC

more data, more power

$N_T = N_B + N_{\text{sig}}$

$N_{\text{sig}} = N_B \times \frac{\sigma_{\text{signal}}}{\sigma_{\text{bkg}}}$

stronger signal
easier to discover

$Z = 5 \sigma$

$N_B = 20\,000$

$Z = 5 \sigma$

A. De Simone
Directions for future work:

- inclusion and impact of uncertainties
- adaptive choice of $K$
- identifying discrepant regions in realistic situations (with $Z$-score method)
- validation tool for bkg: compatibility between MC sims. and data in control regions
- scalability
1. Golden age of Machine Learning
   Big Data are everywhere

2. Innovative Statistical Test for New Physics
   - Powerful and model-independent discovery tool
   - Guidance for experimental searches

3. ML for science in germinal stage
   Pioneering developments waiting ahead!