The ATLAS Higgs Machine Learning Challenge

Claire Adam-Bourdarios¹, Glen Cowan², Cécile Germain-Renaud³, Isabelle Guyon⁴, Balázs, Kégl¹, David Rousseau¹

¹ Laboratoire de l’Accélérateur Linéaire, Orsay, France
² Royal Holloway, University of London, UK
³ Laboratoire de Recherche en Informatique, Orsay, France
⁴ Chalearn, California, USA
Outline

Multivariate analysis in High Energy Physics

The ATLAS Higgs Machine Learning Challenge

https://www.kaggle.com/c/higgs-boson

http://higgsml.lal.in2p3.fr/

C. Adam-Bourdarios et al., Learning to discover: the Higgs boson machine learning challenge, CERN Open Data Portal, DOI: 10.7483/OPENDATA.ATLAS.MQ5J.GHXA

The Problem

The Solutions

Future challenges
Prototype analysis in HEP

Each event yields a collection of numbers \( \vec{x} = (x_1, \ldots, x_n) \)

\[ x_1 = \text{number of muons}, \quad x_2 = p_t \text{ of jet}, \ldots \]

\( \vec{x} \) follows some \( n \)-dimensional joint pdf, which depends on the type of event produced, i.e., signal or background.

1) What kind of decision boundary best separates the two classes?

2) What is optimal test of hypothesis that event sample contains only background?
Machine Learning in HEP

Optimal analysis uses information from all (or in any case many) of the measured quantities → Multivariate Analysis (MVA)

Long history of cut-based analyses, followed by:

1990s    Fisher Discriminants, Neural Networks
Early 2000s  Boosted Decision Trees, Support Vector Machines

But much recent work in Machine Learning only slowly percolating into HEP (deep neural networks, random forests, ...)

Therefore try to promote transmission of ideas from ML into HEP using a Data Challenge.
• Challenges have become in the last 10 years a common way of working for the machine learning community

• Machine learning scientists are eager to test their algorithms on real life problems; more valuable (= publishable) than artificial problems

• Company or academics want to outsource a problem to machine learning scientist, but also geeks, etc. The company sets up a challenge like:
  – Netflix: predict movie preference from past movie selection
  – NASA/JPL mapping dark matter through (simulated) galaxy distortion

• Some companies makes a business from organising challenges: datascience.net, kaggle
The Higgs Machine Learning Challenge
... in a nutshell

• Why not put some ATLAS simulated data on the web and ask data scientists to find the best machine learning algorithm to find the Higgs?
  – Instead of HEP people browsing machine learning papers, coding or downloading a possibly interesting algorithm, trying and seeing whether it can work for our problems

• Challenge for us: make a full ATLAS Higgs analysis simple for non-physicists, but sufficiently close to reality to still be useful for us.

• Also try to foster long-term collaborations between HEP and ML
Committees

• Organization committee:
  – David Rousseau  ATLAS-LAL
  – Claire Adam-Bourdarios  ATLAS-LAL (outreach, legal matters)
  – Glen Cowan  ATLAS-RHUL (statistics)
  – Balázs, Kégl  Appstat-LAL
  – Cécile Germain  TAO-LRI
  – Isabelle Guyon  Chalearn (challenges organization)

• Advisory committee:
  – Andreas Hoecker  ATLAS-CERN (PC,TMVA)
  – Joerg Stelzer  ATLAS-CERN (TMVA)
  – Thorsten Wengler  ATLAS-CERN (ATLAS management)
  – Marc Schoenauer  INRIA (French computer science institute)
Sponsors

This competition is brought to you by

[Logos of various sponsors]

Additional support from:

[More logos of sponsors]
$H \rightarrow \tau^+\tau^-$

4.1 $\sigma$ evidence

Now superseded by
ATLAS paper: Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector, arXiv:1501.04943
Dataset

ASCII csv file, with mixture of Higgs to tautau signal and corresponding background, from official GEANT4 ATLAS simulation

Weight and signal/background (for training dataset only)
weight (fully normalised)
label : « s » or « b »

Conf note variables used for categorization or BDT:

DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality

Primitive 3-vectors allowing to compute the conf note variables (mass neglected),
16 independent variables:

PRI_tau_pt
PRI_tau_eta
PRI_tau_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
PRI_met_sumet
PRI_jet_num (0,1,2,3, capped at 3)
PRI_jet_leading_pt
PRI_jet_leading_eta
PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt
Objective Function

Typical Machine Learning goal is event classification; try to minimize e.g. classification error rate.

Goal in HEP search is to establish whether event sample contains only background; rejecting this hypothesis ≈ discovery of signal.

Often approach in HEP is to use distribution of MVA classifier. Simplest case, use classifier to define “search region” and count:

\[ s = \text{expected number of signal events (assuming it exists)} \]
\[ b = \text{expected number of background events} \]

Goal: Minimize Approximate Median Significance of discovery:

\[ \text{AMS} = \sqrt{2 \left( (s + b) \ln \left( 1 + \frac{s}{b} \right) - s \right)} \]

(Modified in the Challenge to prevent small search region where estimate of \( b \) may fluctuate very low: \( b \rightarrow b + b_{\text{reg.}} \).)

G. Cowan / RHUL Physics

ATLAS Higgs ML Challenge / CHEP 2015
# Real analysis vs challenge

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Systematics</td>
</tr>
<tr>
<td>2.</td>
<td>2 categories x n BDT score bins</td>
</tr>
<tr>
<td>3.</td>
<td>Background estimated from data (embedded, anti tau, control region) and some MC</td>
</tr>
<tr>
<td>4.</td>
<td>Weights include all corrections. Some negative weights (tt)</td>
</tr>
<tr>
<td>5.</td>
<td>Potentially use any information from all 2012 data and MC events</td>
</tr>
<tr>
<td>6.</td>
<td>Few variables fed in two BDT</td>
</tr>
<tr>
<td>7.</td>
<td>Significance from complete fit with NP etc...</td>
</tr>
<tr>
<td>8.</td>
<td>MVA with TMVA BDT</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>No systematics</td>
</tr>
<tr>
<td>2.</td>
<td>No categories, one signal region</td>
</tr>
<tr>
<td>3.</td>
<td>Straight use of ATLAS G4 MC</td>
</tr>
<tr>
<td>4.</td>
<td>Weights only include normalisation and pythia weight. Neg. weight events rejected.</td>
</tr>
<tr>
<td>5.</td>
<td>Only use variables and events preselected by the real analysis</td>
</tr>
<tr>
<td>6.</td>
<td>All BDT variables + categorisation variables + primitives 3-vector</td>
</tr>
<tr>
<td>7.</td>
<td>Significance from “regularised Asimov”</td>
</tr>
<tr>
<td>8.</td>
<td>MVA “no-limit”</td>
</tr>
</tbody>
</table>

**Simpler, but not too simple!**
Participation

• Big success!

• 1785 teams (1942 people) have participated (participation=submission of at least one solution)
  – (6517 people have downloaded the data)
  – most popular challenge on the Kaggle platform, ever (Amazon.com employee access challenge 1687 teams, Allstate Purchase Prediction Challenge 1567 teams)

• 35772 solutions uploaded

• 136 forum topics with 1100 posts
# Final leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>Rank</th>
<th>Team Name</th>
<th>Score</th>
<th>Entries</th>
<th>Last Submission UTC (Best – Last Submission)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>↑1</td>
<td>Gábor Melis ‡ *</td>
<td>$7000</td>
<td>110</td>
<td>Sun, 14 Sep 2014 09:10:04 (-0h)</td>
</tr>
<tr>
<td>2</td>
<td>↑1</td>
<td>Tim Salimans ‡ *</td>
<td>$4000</td>
<td>57</td>
<td>Mon, 15 Sep 2014 23:49:02 (-40.6d)</td>
</tr>
<tr>
<td>3</td>
<td>↑1</td>
<td>nhlxShaze ‡ *</td>
<td>$2000</td>
<td>254</td>
<td>Mon, 15 Sep 2014 16:50:01 (-76.3d)</td>
</tr>
<tr>
<td>4</td>
<td>↑38</td>
<td>ChoKo Team</td>
<td>3.77526</td>
<td>216</td>
<td>Mon, 15 Sep 2014 15:21:36 (-42.1h)</td>
</tr>
<tr>
<td>5</td>
<td>↑35</td>
<td>cheng chen</td>
<td>3.77384</td>
<td>21</td>
<td>Mon, 15 Sep 2014 23:29:29 (-0h)</td>
</tr>
<tr>
<td>6</td>
<td>↑16</td>
<td>quantify</td>
<td>3.77086</td>
<td>8</td>
<td>Mon, 15 Sep 2014 16:12:48 (-7.3h)</td>
</tr>
<tr>
<td>7</td>
<td>↑1</td>
<td>Stanislav Semenov &amp; Co (HSE Yandex)</td>
<td>3.76211</td>
<td>68</td>
<td>Mon, 15 Sep 2014 20:19:03</td>
</tr>
<tr>
<td>8</td>
<td>↓7</td>
<td>Luboš Motl's team ‡ Best physicist</td>
<td>3.76050</td>
<td>589</td>
<td>Mon, 15 Sep 2014 08:38:49 (-1.6h)</td>
</tr>
<tr>
<td>9</td>
<td>↑18</td>
<td>Roberto-UCLIMM</td>
<td>3.75864</td>
<td>292</td>
<td>Mon, 15 Sep 2014 23:44:42 (-44d)</td>
</tr>
<tr>
<td>10</td>
<td>↑12</td>
<td>Davut &amp; Josef ‡</td>
<td>3.75838</td>
<td>161</td>
<td>Mon, 15 Sep 2014 23:24:32 (-4.5d)</td>
</tr>
<tr>
<td>45</td>
<td>↑15</td>
<td>crowwork ‡ HEP meets ML award XGBoost authors Free trip to CERN</td>
<td>3.71885</td>
<td>94</td>
<td>Mon, 15 Sep 2014 23:45:00 (-5.1d)</td>
</tr>
<tr>
<td>782</td>
<td>↓149</td>
<td>Eckhard TMVA expert, with TMVA improvements</td>
<td>3.49945</td>
<td>29</td>
<td>Mon, 15 Sep 2014 07:26:13 (-46.1h)</td>
</tr>
</tbody>
</table>

G. Cowan / RHUL Physics
Gabor Private(#1) = 3.806  Public(#2) = 3.786

% rejected

AMS

significance

Public

Private
Lubos Private(#7) = 3.760  Public(#1) = 3.851

Clear overtraining!
Private leaderboard top AMS curves

AMS

% rejected

Lubos
ChoKo
Gabor
quantify
Roberto
Pierre
multiboost

Clearly at the top!
What did we learn

• Very successful satellite workshop at NIPS in Dec 2014 @ Montreal:  [https://indico.lal.in2p3.fr/event/2632/](https://indico.lal.in2p3.fr/event/2632/)

20% gain w.r.t. to untuned TMVA

**Deep Neural nets** (but marginally better than BDT)

**Ensemble methods** (random forest, boosting) rule

**Meta-ensembles** of diverse models

**careful cross-validation** (250k training sample really small)

Complex software suites using routinely multithreading, GPU, etc…

Some techniques (e.g. Meta-ensembles) too complex to be practical, and
marginal gain, others appear practical and useful
Next steps

Re-importing into HEP all the ML developments

Dataset will remain on CERN Open Data Portal with citeable d.o.i.:
http://opendata.cern.ch/education/ATLAS

- Release with the full truth info

Better understand what was done by the best participants

NIPS proceedings write-up (with description of “how they did it”)

Organisation of visit of winners of HEP meets ML award at CERN
(authors of XGBoost Tianqi Chen and Tong He, and overall winner Gabor Melis)

Mini workshop 19th May 2015 2PM in CERN Auditorium,
http://cern.ch/higgsml-visit

Discussion on-going with TMVA experts
Extra slides
Ol is suitable for a variety of nonconventional surprising ideas that are « far » from traditional expertise - > high volatility

Experts are highly skilled, trained - > more focused, performed solution, low variety

Not just ML, but a general trend: Open Innovation

Why challenges work?

MOTIVATION OF ORGANIZING CONTESTS: EXTREME VALUE

Courtesy: Lakhani 2014
From domain to challenge and back

Domain e.g. HEP

Problem

Domain experts solve the domain problem

Solution

Challenge

Problem

simplify

The crowd solves the challenge problem

Solution

reimport