Incorporation of Systematic Uncertainties in the Training of Multivariate Methods

Thomas Alef & Eckhard von Törne, University of Bonn

ACAT 2019
March 13th, 2019
Motivation and Goal

- Modern analysis often limited by systematic uncertainties ⇒ make multivariate methods robust against systematics
- **Systematic aware Boosted Decision Trees (saBDT)** developed during Masterthesis
  - Based on AdaBoost/Gini Index BDTs from TMVA
  - Tested on modified public data from Kaggle Higgs Challenge
- Compared with **Adversarial Neural Networks (AdvNN)**
- **AdvNN** based on KERAS
Public Data from Kaggle Higgs Challenge

- Data from Kaggle Higgs Challenge $H \rightarrow \tau\tau$
- 30 variables
- Training: 120,000 events (Kaggle challenge public data)
- Evaluation: 550,000 events (Kaggle challenge private data)
- For testing influence of systematics a systematic variation was added
Implementation of Systematic Variation

- Jet Energy Scale chosen as example systematic - standard ATLAS systematic
- Strength of systematic variation: 1% (ATLAS standard value 1-4%)
- Scale jet energies up by 1%
  → recalculate all variables based on jet energies with new values
  → new systematic varied *Up* dataset
- Repeat with scaling down by 1%
- ⇒ 3 Datasets: *Nominal*, *Up*, *Down*

![Diagram showing training and evaluation datasets]
Evaluation Metric: AAMS

- Kaggle Challenge used Approximate Median Significance (AMS)
- Adding systematic uncertainty: Advanced Approximate Median Significance (AAMS) (see hal-01208587)
- **Cut and Count** approach: events with higher score than $x$ are classified signal

\[
AAMS = \sqrt{2 \left( \frac{(s + b) \ln s + b}{b_0} - s - b + b_0 \right) + \frac{(b - b_0)^2}{\sigma_b^2}}
\]

\[
b_0 = \frac{1}{2} \left( b - \sigma_b^2 + \sqrt{(b - \sigma_b^2)^2 + 4(s + b)\sigma_b^2} \right)
\]

- $s$ signal events, $b$ background events, $\sigma_b$ background difference on the different data sets
- Unstable for small $b$ $\rightarrow$ add a regularization term of 10 to $b$
- Maximum of $\sigma_b$ for all possible cut values: $\sigma_b^{\text{max}}$
  $\Rightarrow$ if small, method behaves similar on varied datasets
How to make BDTs aware of Systematics

BDT uses all three datasets during training
If performance similar on all three datasets - invariant under systematic variations
Similar behavior checked for:
  - Every single node split
  - Whole tree (Boostweight)
AdaBoost BDT with Gini Index on ROOT 6.10/06
NTrees=1000, MinNodeSize=1%, AdaBoost=0.2

BDTOut = Weight(Tree1) * Result (Tree1) + Weight(Tree2) * Result (Tree2) + ...
BDT: Standard Node Split

- So far: scan through all variables and possible cuts, maximize:

  \[ \text{Gain} = G_{\text{Parent}} - G_{\text{Left}} - G_{\text{Right}} \]

  with Gini Index \( G = p \cdot (1 - p) \) (maximal for \( p=0.5 \)) and \( p = \frac{N_{\text{Signal}}}{N_{\text{All}}} \)

- Basically: find the cut which improves the purity of the nodes the most
saBDT: Systematic Aware Node Split

- Modify $Gain$ to penalize differing behavior on different data sets
- Modification based on purity to stay consistent
- Subtract a term accounting for purity differences on different data sets:

$$NewGain = Gain - \lambda_{Cut} \cdot \frac{1}{8} \cdot \sqrt{\sum_{Left,Right} (p_{Reg} - p_{Up,Down})^2}$$

- $\lambda_{Cut}$ as hyperparameter to control strength of invariance
- Penalty term can be between 0 and 0.25
saBDT: $\lambda_{\text{Cut}}$ Hyperparameter Scan Results

- Stable AAMS with possible increase for low $\lambda_{\text{Cut}}$
  $\Rightarrow$ Algorithm works!
- $\sigma^\text{max}_b$ decreases - performance similar on different datasets
Every decision tree is weighted according to its error rate:

$$\text{err} = \frac{N_{\text{misidentified}}}{N_{\text{All}}} \Rightarrow TW = \log\frac{1 + \text{err}}{1 - \text{err}}$$

- $TW$ is the boost weight, high when tree performing well
- Multiply factor accounting for differences on systematic varied samples:

$$\text{New } TW = TW \cdot \exp\left(-\lambda_{\text{Boost}} \cdot \sum_{\text{Up,Down}} \frac{(err_{\text{Reg}} - err_{\text{Up,Down}})^2}{2}\right)$$

- $\lambda_{\text{Boost}}$ as hyperparameter
- New factor pulls down weight of trees affected by systematic variation
saBDT: $\lambda_{\text{Boost}}$ Hyperparameter Scan Results

- AAMS (performance) drops for high $\lambda_{\text{Boost}}$
- Influence of systematics decreases as well!
- Stable region with possible increase for low values

![Graph showing AAMS vs $\lambda_{\text{Boost}}$]
saBDTs: 2-Dim Hyperparameter Scan

- Scanning through $\lambda_{\text{Cut}}$ and $\lambda_{\text{Boost}}$ reveals increasing AAMS
- Confirmed by Bootstrap: 82.1% chance it is not a statistical fluctuation
saBDT: Different Strength of Systematic Variation

- Different strength in systematic variation of data is applied

<table>
<thead>
<tr>
<th>Systematic Variation</th>
<th>BDT (AAMS)</th>
<th>saBDT (AAMS)</th>
<th>% no stat. Fluc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>1.07±0.05</td>
<td>1.52±0.06</td>
<td>98.4%</td>
</tr>
<tr>
<td>10%</td>
<td>1.38±0.06</td>
<td>1.94±0.07</td>
<td>99.6%</td>
</tr>
<tr>
<td>3%</td>
<td>2.40±0.09</td>
<td>2.64±0.09</td>
<td>92.3%</td>
</tr>
<tr>
<td>1%</td>
<td>3.13±0.11</td>
<td>3.22±0.10</td>
<td>82.1%</td>
</tr>
</tbody>
</table>

- 3% and 1% ATLAS standard values
- saBDTs improves result especially well with strong systematic variation
- Result dominated by systematic uncertainty in this region → decreasing systematic uncertainty more valuable
Adversarial Neural Networks

- As comparison AdvNN (see Louppe, Kagan, Cranmer: arXiv:1611.01046)
- Multiple talks during the next days
- Classifier able to distinguish signal and background
- Adversary penalizing Classifier if it is sensitive to systematic variations
- $\gamma$ as strength parameter for penalty
saBDTs vs AdvNNs

- Comparison for 1% systematic variation
- saBDT performs slightly better!
- Maximal $AAMS(\text{saBDT}) = 3.23 \pm 0.10$, $AAMS(\text{AdvNN}) = 3.08 \pm 0.11$
- AdvNN not fully optimized
Conclusion

- saBDTs proved capable of reducing systematic uncertainty
- Gain in AAMS was achieved
- AdvNNs were outperformed
- AdvNNs less optimized than saBDTs - difference originating from this?
- Invariance proved to be most valuable for high systematic effects

Outlook

- saBDTs tested with different systematics
- New metrics to test the performance
- Multiple systematics at once?
Backup
saBDT: Node Split BDT Distribution

\[ \lambda_{\text{Cut}} = 0 \]

\[ \lambda_{\text{Cut}} = 0.001 \]

\[ \lambda_{\text{Cut}} = 0.02 \]
Distributions behave similar to $\lambda_{\text{Cut}}$

Getting shifted to the left
saBDT: AAMS

Significance

Dataset
- Standard
- $\lambda_{\text{Cut}}=0.001$, $\lambda_{\text{Boost}}=2$

AAMS

Cut

-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8
Adversarial Neural Networks

Used AvdNN

- **Classifier**:
  - 30 input nodes, one for every variable
  - 3 dense hidden layers, regularized by $l_1 = 0.0001$ and $l_2 = 0.001$
  - 120 nodes each
  - Activation function is *relu* for the hidden layers
  - 1 output note, with *sigmoid* as activation function
  - batch size is 64

- **Adversary**:
  - 1 input node
  - 3 dense hidden layers
  - The first two hidden layers have 30 nodes each and the last with 12
  - Activation function is *relu* for the hidden layers
  - 3 output nodes, with *softmax* as activation
# Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>DER_mass_MMC</td>
<td>effect but hard to calculate – neglected</td>
</tr>
<tr>
<td>DER_mass_transverse_met_lep</td>
<td>If mEt is affected here as well</td>
</tr>
<tr>
<td>DER_mass_vis</td>
<td>not affected</td>
</tr>
<tr>
<td>DER_pt_h</td>
<td>If mEt is affected here as well</td>
</tr>
<tr>
<td>DER_deltaeta_jet_jet</td>
<td>not affected</td>
</tr>
<tr>
<td>DER_mass_jet_jet</td>
<td>directly affected</td>
</tr>
<tr>
<td>DER_prodeta_jet_jet</td>
<td>not affected</td>
</tr>
<tr>
<td>DER_deltar_tau_lep</td>
<td>not affected</td>
</tr>
<tr>
<td>DER_pt_tot</td>
<td>directly affected</td>
</tr>
<tr>
<td>DER_sum_pt</td>
<td>directly affected</td>
</tr>
<tr>
<td>DER_pt_ratio_lep_tau</td>
<td>not affected</td>
</tr>
<tr>
<td>DER_met_phi_centralty</td>
<td>If mEt is affected here as well</td>
</tr>
<tr>
<td>DER_lep_eta_centralty</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_tau_pt</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_tau_eta</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_tau_phi</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_lep_pt</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_lep_eta</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_lep_phi</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_met</td>
<td>affected similar to jet energy</td>
</tr>
<tr>
<td>PRI_met_phi</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_met_sumet</td>
<td>directly affected</td>
</tr>
<tr>
<td>PRI_jet_num</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_jet_leading_pt</td>
<td>directly affected</td>
</tr>
<tr>
<td>PRI_jet_leading_eta</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_jet_leading_phi</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_jet_subleading_pt</td>
<td>directly affected</td>
</tr>
<tr>
<td>PRI_jet_subleading_eta</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_jet_subleading_phi</td>
<td>not affected</td>
</tr>
<tr>
<td>PRI_jet_all_pt</td>
<td>directly affected</td>
</tr>
</tbody>
</table>
saBDT: systematic aware node split results

- AAMS (performance) drops initially with $\lambda_{\text{Cut}}$
- Influence of systematics decreases as well!
- Breakdown around $\lambda_{\text{Cut}} = 0.01$

![Graph showing AAMS and $\sigma_B^{\text{max}}$ vs. $\lambda_{\text{Cut}}$](image)
saBDT: $\lambda_{Boost}$ Hyperparameter Scan Results

- **AAMS** (performance) drops for high $\lambda_{Boost}$
- Influence of systematics decreases as well!
- Stable region with possible increase for low values

\[ AAMS \]

\[ \sigma_{B}^{\text{max}} \]
Overview of AMS/AAMS Results

<table>
<thead>
<tr>
<th>Method</th>
<th>AMS</th>
<th>AAMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaggle Winner</td>
<td>3.81</td>
<td>NA</td>
</tr>
<tr>
<td>Kaggle TMVA</td>
<td>3.50</td>
<td>NA</td>
</tr>
<tr>
<td>BDT</td>
<td>3.44</td>
<td>3.13</td>
</tr>
<tr>
<td>saBDT</td>
<td>3.35</td>
<td>3.22</td>
</tr>
<tr>
<td>NN</td>
<td>3.27</td>
<td>2.88</td>
</tr>
<tr>
<td>AdvNN</td>
<td>3.20</td>
<td>3.08</td>
</tr>
</tbody>
</table>

- Including systematic aware training leads to loss in \textit{AMS} and gain in \textit{AAMS}
- Tested methods not as fully optimized as during challenge
Aware Boosted Decision Trees: Bootstrap

- Difference in performance of standard BDT and tuned saBDT tested on bootstrapped samples
- Bootstrap creates new samples with different statistics out of the original sample
- saBDT performs indeed better, but not significant
- $\Delta AAMS = 0.138 \pm 0.150$