CS460 - Machine Learning 2023

INSTRUCTOR: Dr. Subhankar Mishra

Project Proposal

Gradient Boosted Decision Trees and their application in improvising identification of decay of Higgs Bosons into pair of electrons.

Team members:

- Arpan Maity
- Krishna Kant Parida

Proposed Project timeline

Dataset Evaluation:

- Monte Carlo Generated Data [Training, Validation]
- Real time p-p collision data (reconstructed) at $\sqrt{s} = 13$ TeV and IL = 138 fb^{-1} [Testing]

↓

Model training and Hyperparameter tuning and comparison with other models (e.g. AdaBoost)

↓

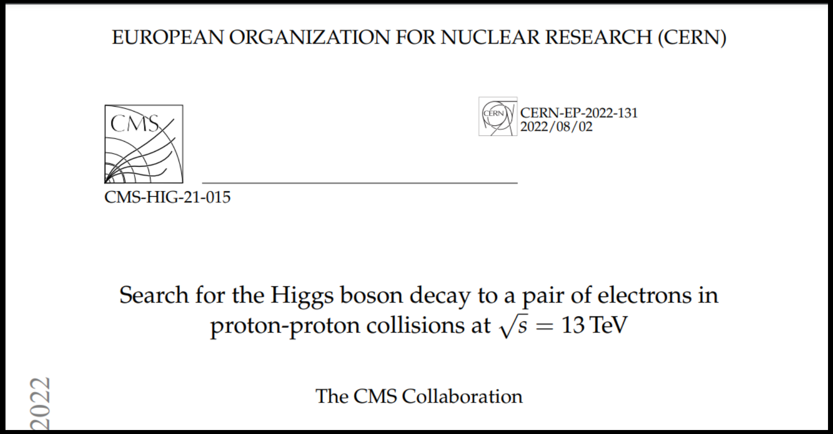
Testing and further improvisations (stochastic based modelling etc.)

Proposed Work Division

- Division on evaluating datasets (Real time dataset Provided By Prof. Sanjay Swain)
- Division on Training and hyperparameter tuning
- Testing and improvisation (TBD)

Base Paper

1. CMS Collaboration, Search for the Higgs boson decay to a pair of electrons in proton-proton collisions at √s = 13 TeV," CERN-EP-2022-131, CMS-HIG-21-015
2. T. Hastie, R. Tibshirani, J. Friedman "Elements of statistical learning" 2nd Ed. Ch-10, 10.10.2 Page 359



EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

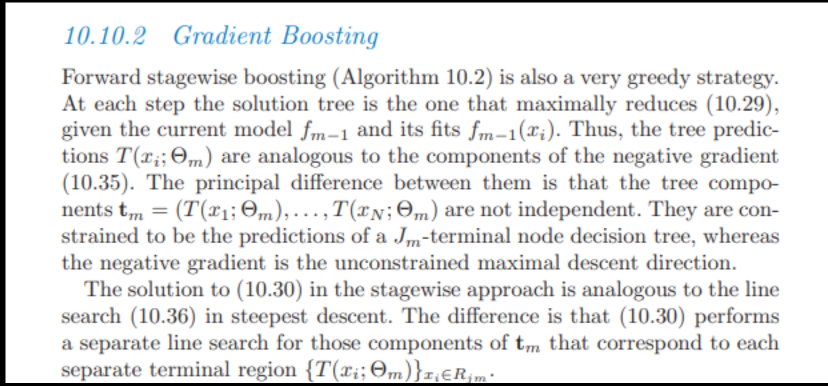
CERN-EP-2022-131
2022/08/02

CMS-HIG-21-015

Search for the Higgs boson decay to a pair of electrons in proton-proton collisions at $\sqrt{s} = 13$ TeV

The CMS Collaboration

2022



10.10.2 Gradient Boosting

Forward stagewise boosting (Algorithm 10.2) is also a very greedy strategy. At each step the solution tree is the one that maximally reduces (10.29), given the current model f_{m-1} and its fits $f_{m-1}(x_i)$. Thus, the tree predictions $T(x_i; \Theta_m)$ are analogous to the components of the negative gradient (10.35). The principal difference between them is that the tree components $\mathbf{t}_m = (T(x_1; \Theta_m), \ldots, T(x_N; \Theta_m))$ are not independent. They are constrained to be the predictions of a J_m-terminal node decision tree, whereas the negative gradient is the unconstrained maximal descent direction.

The solution to (10.30) in the stagewise approach is analogous to the line search (10.36) in steepest descent. The difference is that (10.30) performs a separate line search for those components of $\mathbf{t}_m$ that correspond to each separate terminal region $\{T(x_i; \Theta_m)\}_{x_i \in R_{jm}}$.

THANK YOU