Queen Mary
University of London

DR ADRIAN BEVAN

PRACTICAL MACHINE LEARNING

$H \rightarrow \tau^{+}\tau^{-}$

LECTURE PLAN

- Introduction
- The problem
- The features
- The data
- Starting example
- Summary

QMUL Summer School: https://www.qmul.ac.uk/summer-school/
Practical Machine Learning QMplus Page: https://qmplus.qmul.ac.uk/course/view.php?id=10006

INTRODUCTION

- The Higgs boson is an important part of the Standard Model of particle physics (SM).
- We need to verify that the particle discovered in 2012 behaves in the same way that the Higgs boson is expected to for the SM.
  - It is possible that we have found a particle that looks like the Higgs, but is different in some way.
  - The Higgs was invented to solve a problem with electroweak symmetry breaking: the mass splitting between the W and Z bosons.
  - For convenience the same particle is assumed to lead to mass generation for fermions (quarks and leptons).
  - Important to measure the particle decaying into different types of decay products: including bosons, quark and lepton pairs to validate that the Higgs boson behaves as we expect for the SM.

INTRODUCTION

- The decay probability for a Higgs particle to pairs of fermions is

Decay channel	Probability (%)
H→bb	57.7
H→WW	21.5
H→ττ	6.3
H→ZZ	2.6
H→γγ	0.2

- In addition to signal, there are significant background channels that mean that the the best measured channels are for γγ and ZZ final states.

- The $H \to \tau^+\tau^-$ channel is an important decay to measure, and this requires separation of signal from background.

INTRODUCTION

- The ATLAS experiment at CERN released some data and Monte Carlo simulated data via a Kaggle data challenge.
- We will be using these samples for the Week 3 assignment.
- The Higgs Kaggle Web page can be found at:
  - https://www.kaggle.com/c/higgs-boson
- Documentation can also be found on the preprint archive:
  - https://higgsml.lal.in2p3.fr

THE PROBLEM

- Develop a model that can be used to distinguish between signal and background for the $H \to \tau^+\tau^-$ sample.
- There are several simplifications for this task relative to a normal HEP analysis:
  - Events with negative weights have been removed (comes from Monte Carlo generators of some simulators).
  - Only the dominant background sources are included.
  - Some correction factors have been neglected.

THE PROBLEM

- The approximate median significance is a metric used to compare results. This is given by

$$\mathrm{AMS} = \sqrt{2\left((s+b+b_{\mathrm{reg}})\ln\left(1+\frac{s}{b+b_{\mathrm{reg}}}\right)-s\right)}$$

- The term b_{reg} is used to stop the search reverting to small regions of the feature space where statistical fluctuations can become significant. This is set to 10 for the challenge.
- s and b:
  - are defined in the challenge notes as the sum over the example weights for the signal and background events used in the search region.
  - They are unbiased estimators of the number of signal and background events, respectively.

(see sec. 2 of the challenge notes)

THE FEATURES

- Some formulae (physics) are included in Appendix A of "Learning to discover: the Higgs boson machine learning challenge" for the interested student (context).
- Features are listed in Appendix B.
- The following features are **NOT** to be used in the classifier:
  - EventId: A unique integer identifier of the example[1].
  - Weight: Event weight.
  - Label[2]: The event label (string) yi 2 {s, b} (s for signal, b for background).
  - KaggleSet: Specific to the opendata.cern.ch dataset: string specifying to which Kaggle set the event belongs: "t":training, "b":public leaderboard, "v":private leaderboard, "u":unused.
  - KaggleWeight: Specific to the opendata.cern.ch dataset: weight normalized within each Kaggle data set according to:

(see Appendix B of the challenge notes)

$$w'_j = w_j \frac{\sum_i w_i \mathbb{1}\{y_i = y_j\}}{\sum_{i \in S'} w_i \mathbb{1}\{y_i = y_j\}}$$

[1] In HEP training examples are normally referred to as events, as is the case in the documentation associated with this challenge.
[2]Not available in the test sample.

THE FEATURES

- Features listed from here on are usable in the classifier.
- Features with names prefixed with PRI and DER are:
  - PRI: Primary features “raw“ quantities measured on objects like jets.
  - DER: Derived features - combinations of the primary features derived from lower level (raw) information.

(see Appendix B of the challenge notes)

THE FEATURES

DER_mass_MMC The estimated mass m_H of the Higgs boson candidate, obtained through a probabilistic phase space integration (may be undefined if the topology of the event is too far from the expected topology)

DER_mass_transverse_met_lep The transverse mass (22) between the missing transverse energy and the lepton.

DER_mass_vis The invariant mass (21) of the hadronic tau and the lepton.

DER_pt_h The modulus (20) of the vector sum of the transverse momentum of the hadronic tau, the lepton, and the missing transverse energy vector.

DER_deltaeta_jet_jet The absolute value of the pseudorapidity separation (23) between the two jets (undefined if `PRI_jet_num` ≤ 1).

DER_mass_jet_jet The invariant mass (21) of the two jets (undefined if `PRI_jet_num` ≤ 1).

DER_prodeta_jet_jet The product of the pseudorapidities of the two jets (undefined if `PRI_jet_num` ≤ 1).

DER_deltar_tau_lep The R separation (24) between the hadronic tau and the lepton.

DER_pt_tot The modulus (20) of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet (if `PRI_jet_num` ≥ 1) and the subleading jet (if `PRI_jet_num` $= 2$) (but not of any additional jets).

DER_sum_pt The sum of the moduli (20) of the transverse momenta of the hadronic tau, the lepton, the leading jet (if `PRI_jet_num` ≥ 1) and the subleading jet (if `PRI_jet_num` $= 2$) and the other jets (if `PRI_jet_num` $= 3$).

DER_pt_ratio_lep_tau The ratio of the transverse momenta of the lepton and the hadronic tau.

(see Appendix B of the challenge notes)

THE FEATURES

DER_met_phi_centrality The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton

$$C = \frac{A + B}{\sqrt{A^2 + B^2}},$$

where $A = \sin(\phi_{\text{met}} - \phi_{\text{lep}}) * \text{sign}(\sin(\phi_{\text{had}} - \phi_{\text{lep}}))$, $B = \sin(\phi_{\text{had}} - \phi_{\text{met}}) * \text{sign}(\sin(\phi_{\text{had}} - \phi_{\text{lep}}))$, and ϕ_{met}, ϕ_{lep}, and ϕ_{had} are the azimuthal angles of the missing transverse energy vector, the lepton, and the hadronic tau, respectively. The centrality is $\sqrt{2}$ if the missing transverse energy vector $\vec{E}_{\text{T}}^{\text{miss}}$ is on the bisector of the transverse momenta of the lepton and the hadronic tau. It decreases to 1 if $\vec{E}_{\text{T}}^{\text{miss}}$ is collinear with one of these vectors and it decreases further to $-\sqrt{2}$ when $\vec{E}_{\text{T}}^{\text{miss}}$ is exactly opposite to the bisector.

DER_lep_eta_centrality The centrality of the pseudorapidity of the lepton w.r.t. the two jets (undefined if `PRI_jet_num` ≤ 1)

$$\exp\left[\frac{-4}{(\eta_1 - \eta_2)^2}\left(\eta_{\text{lep}} - \frac{\eta_1 + \eta_2}{2}\right)^2\right],$$

where η_{lep} is the pseudorapidity of the lepton and η_1 and η_2 are the pseudorapidities of the two jets. The centrality is 1 when the lepton is on the bisector of the two jets, decreases to $1/e$ when it is collinear to one of the jets, and decreases further to zero at infinity.

PRI_tau_pt The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the hadronic tau.

PRI_tau_eta The pseudorapidity η of the hadronic tau.

PRI_tau_phi The azimuth angle ϕ of the hadronic tau.

PRI_lep_pt The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the lepton (electron or muon).

(see Appendix B of the challenge notes)

Queen Mary University of London

THE FEATURES

PRI_lep_eta The pseudorapidity η of the lepton.

PRI_lep_phi The azimuth angle ϕ of the lepton.

PRI_met The missing transverse energy $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$.

PRI_met_phi The azimuth angle ϕ of the missing transverse energy.

PRI_met_sumet The total transverse energy in the detector.

PRI_jet_num The number of jets (integer with value of 0, 1, 2 or 3; possible larger values have been capped at 3).

PRI_jet_leading_pt The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the leading jet, that is the jet with largest transverse momentum (undefined if `PRI_jet_num` $= 0$).

PRI_jet_leading_eta The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` $= 0$).

PRI_jet_leading_phi The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` $= 0$).

PRI_jet_subleading_pt The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the leading jet, that is, the jet with second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1).

PRI_jet_subleading_eta The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` ≤ 1).

PRI_jet_subleading_phi The azimuth angle ϕ of the subleading jet (undefined if `PRI_jet_num` ≤ 1).

PRI_jet_all_pt The scalar sum of the transverse momentum of all the jets of the events.

(see Appendix B of the challenge notes)

THE DATA

- The following data samples are available:
  - All of the data: atlas-higgs-challenge-2014-v2.csv
    (818239 examples: signal, background and all KaggleLabel types)
  - Training data: train_*.csv
    (85668 signal and 164334 background examples)
  - Test data: test_*.csv
    (34026 signal and 65976 background examples)
  - Unused data: unused_*.csv
    (6186 signal and 12054 background examples)
  - Private test data[1]: train_private_*.csv
    (153684 signal and 296318 background examples)
  - Training and test data are also split into signal and background for convenience (i.e. * = sig, bg).

[1]This sample was reserved for leaderboard validation of the test.

THE DATA

- The following data samples are available:
  - All of the data: atlas-higgs-challenge-2014-v2.csv
    (818239 examples: signal, background and all KaggleLabel types)
  - Training data: train_*.csv
    (85668 signal and 164334 background ex
  - Test data: test_*.csv
    (34026 signal and 65976 background exa
  - Unused data: unused_*.csv
    (6186 signal and 12054 background examples)
  - Private test data[1]: train_private_*.csv
    (153684 signal and 296318 background examples)
  - Training and test data are also split into signal and background for convenience (i.e. * = sig, bg).

Two small training files with 5k examples have been provided as train_sml_(sig/bg).csv to be used when debugging code and setting up model options.

[1]This sample was reserved for leaderboard validation of the test.

STARTING EXAMPLE

- `Example_KaggleHiggs.py`

- The example is set up to read in a training sample (the small sample of 5K sig/bg events).
- A merged data set totalling 1000 events is then constructed.
- These are run through training for a single layer perceptron with 256 nodes in the hidden layer.
- The training cost and accuracy as a function of epoch is recorded for plotting.
- This should work out of the box and provide a baseline reference.

STARTING EXAMPLE

- `Example_KaggleHiggs.py`

- Parameters used to configure the training can be found at the top of the script under '`global network configuration`'
- `RunAnalysis` is the function that is called to train and test the model.
- The model is implemented following the comment starting with '`setup the model`'
- The training is run following the comment starting with '`train the model`'
- The function '`BuildFeatureSpace`' controls what features are included as inputs to the network. By default only the following are used:
  - DER_mass_MMC, DER_mass_transverse_met_lep, DER_mass_vis, DER_pt_tot, DER_sum_pt, PRI_tau_pt, PRI_lep_pt, PRI_met
- You will want to experiment with the features used in the network as well as the network model and training parameters.

STARTING EXAMPLE

- `Example_KaggleHiggs.py`

- This starting example has been provided for you to work from.
  - Use your knowledge from the previous weeks to work toward a model for this problem.
  - Check for over training using your test sample.
  - Work toward the best accuracy that you can.
- As you develop your model and validation code you will want to change the data sample used to train from the small sample to the nominal training sample.
- When submitting your portfolio of work for this, provide a summary of model configurations that you have tried, with accuracies and where possible the AMS values obtained.
  - Remember - do not submit the data files as well as your code etc. If you try that you will not be able to submit your portfolio of work for this week!

SUMMARY

- This is the last week of the course, and the work to be done this week involves making the most accurate model for separating signal from background, building on the previous two weeks worth of work.
- Results and code should be submitted for assessment at the end of the week.
- Those of you taking this course for credit will also have an oral presentation to complete at the end of the week.
  - Plan your time accordingly to make sure you meet submission deadlines.