Higgs Physics

Exercise Sheet 10

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Fall semester 2015
Lecturers: M. Donega, M. Grazzini
Assistants: P. Musella, H. Sargsyan
https://moodle-app2.let.ethz.ch/course/view.php?id=1720

Issued : 23.11.2015
Due : 27.11.2015

Exercise 1 [*Event classification using decision trees.*]

This exercise illustrates the use of the decision trees for event classification. The problem that will be treated is the discrimination between the diphoton decay of the Higgs boson in VBF production at the LHC and the irreducible background from QCD production of two photons.

You will need a few software tools:

- The CERN ROOT package https://root.cern.ch/downloading-root
- The sci-kit framework http://scikit-learn.org/stable/install.html
- The jupyter notebook http://jupyter.readthedocs.org/en/latest/install.html
- A few dependencies: root_numpy and matplotlib

In practice, after installing ROOT, you can obtain the other packages as

```
conda install numpy scipy scikit-learn jupyter matplotlib root_numpy
```

or

```
pip install numpy scipy scikit-learn jupyter matplotlib root_numpy
```

a) A few examples of how to train decision trees using the sci-kit package are provided.

- Download the files needed for the exercises, *Decision trees.ipynb*, *signal.root* and *background.root*.
- The ROOT files contain VBF production of a Higgs boson decaying to two photons and QCD production of two photons respectively. Only events with at least two jets of transverse momentum above 20 GeV are saved.
- Several variables are saved in the trees.
  - *mass, pt, rapidity and eta* are the characteristics of the diphoton system.
  - *leadPt, leadEta, leadPhi* and the analogous *sublead* variables refer to the higher and lower pt photon.
  - *leadJetPt, leadJetEta* and the analogous *sublead* variables refer to the higher and lower pt jets.
  - *cosDphiJJ, deltaEtaJJ and mjj* represent the cosine of azimutal angle between the two jets, the absolute value of the rapidity difference and the invariant mass of the dijet system respectively.
  - *zepVar* is defined as the difference between the pseudo-rapidity of the diphoton system and the average of the pseudo rapidity of the two jets.
  - *cosDphiGGJJ* the cosine of azimutal angle between the diphoton and the dijet system.

b) The examples provided in the notebook focus on three of the input variables, *pt, deltaEtaJJ* and *mjj*.

- Do you expect these variables to be correlated? If so, do you expect that the correlation would be different for signal and background?
- The input dataset is first split into two different subsets, labelled *train* and *test*. What is the purpose of this procedure?
- In the notebook, decision trees of different depths are built using pairs of input variables. Are deeper trees improving the separation between signal and background?
- The same exercises is carried out for decision trees using all the three variables. A quantitative measure of the goodness of the trained trees is provided histogramming the output of the decision tree on signal and background events. Looking at these plots, what do you conclude on the use deeper decision trees?

c) The use of adaptive boosting is shown at the end of the notebook.

- How does the performance of the boosted trees compare with the one of the simple trees?
- The performance of the BDT is evaluated looking at the so called "Receiver Operating Characteristics", which is constructed plotting the background vs signal efficiency. How would a "perfect" discriminant appear on such a plane? And one which does not distinguish the two classes?

d) Now try to extend the list of input variables in order to improve the discrimination between signal and background.

- Include *zepVar, cosDphiJJ, leadJetPt and subleadJetPt* in the list of input variables.
- Make 1D histograms of each of the variables for signal and background events. Are these variables useful to discriminate between the two processes?
- Train a boosted decision tree using these new variables on top of the three of the previous example.
- Use the ROC curve in order to quantify the improvement brought by the addition of these variables.