Computational methods applied to Particle Physics

Higgs Boson Machine Learning Challenge

Z. Kassabov

July 2014

Contents

Contents

Discovery of the Higgs boson

In 2013 ATLAS and CMS experiments confirmed that the boson found in 2012 is compatible with a Higgs boson.

- Mass of 125 GeV.
- Spin 0.
- Decays $\gamma\gamma$, ZZ and W^+W^- as predicted by SM.
  - The Brout–Englert–Higgs mechanism is responsible for the mass of the W and Z bosons.

Discovery of the Higgs boson

In 2013 ATLAS and CMS experiments confirmed that the boson found in 2012 is compatible with a Higgs boson.

- Mass of 125 GeV.
- Spin 0.
- Decays $\gamma\gamma$, ZZ and W^+W^- as predicted by SM.
  - The Brout–Englert–Higgs mechanism is responsible for the mass of the W and Z bosons.

Discovery of the Higgs boson

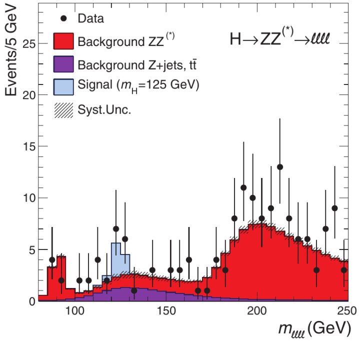
In 2013 ATLAS and CMS experiments confirmed that the boson found in 2012 is compatible with a Higgs boson.

- Mass of 125 GeV.
- Spin 0.
- Decays $\gamma\gamma$, ZZ and W^+W^- as predicted by SM.
    - The Brout–Englert–Higgs mechanism is responsible for the mass of the W and Z bosons.

Discovery of the Higgs boson

In 2013 ATLAS and CMS experiments confirmed that the boson found in 2012 is compatible with a Higgs boson.

- Mass of 125 GeV.
- Spin 0.
- Decays $\gamma\gamma$, ZZ and W^+W^- as predicted by SM.
    - The Brout–Englert–Higgs mechanism is responsible for the mass of the W and Z bosons.



Fermion couplings

The SM Higgs couples to fermions, via Yukawa couplings.

- Branching ratios predicted to scale with the mass squared of the decay products.
    - Most abundant fermionic decay channels:

$$H \to \tau^+\tau^-$$

$$H \to b\bar{b}$$

- Some theories predict a different mechanism for the masses of the leptons (eg, more than one Higgs field).

Fermion couplings

The SM Higgs couples to fermions, via Yukawa couplings.

- Branching ratios predicted to scale with the mass squared of the decay products.
  - Most abundant fermionic decay channels:

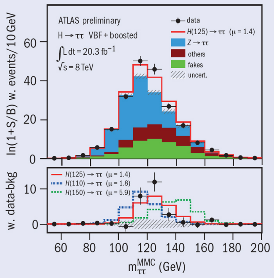
$$H \to \tau^+\tau^-$$

$$H \to b\bar{b}$$

- Some theories predict a different mechanism for the masses of the leptons (eg, more than one Higgs field).

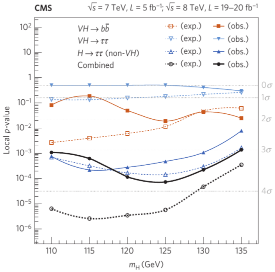
Evidence for fermion couplings

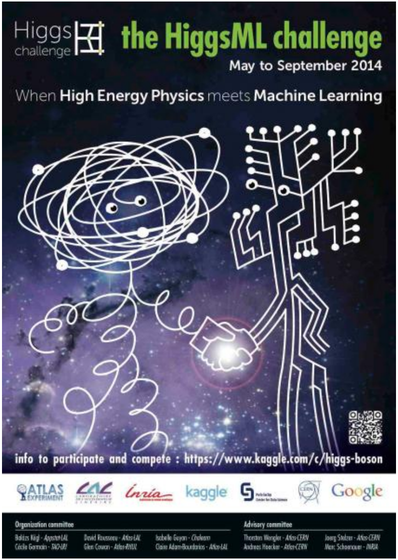
- 2013, ATLAS experiment.
  - Significance of 4.1σ in the $\tau\tau$ channel.
- June 2014, CMS experiment.
  - Significance of 3.4σ in the $\tau\tau$ channel.
  - Significance of 4.0σ combined with the bb channel.



Evidence for fermion couplings

- 2013, ATLAS experiment.
  - Significance of 4.1σ in the $\tau\tau$ channel.
- June 2014, CMS experiment.
  - Significance of 3.4σ in the $\tau\tau$ channel.
  - Significance of 4.0σ combined with the bb channel.





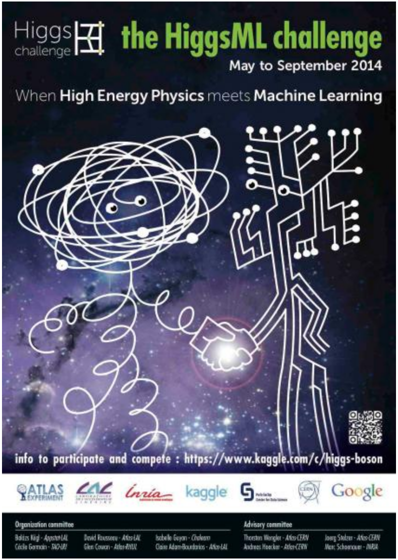
The goal of the challenge is to explore the use of Machine Learning tools to improve the discovery significance of the experiment.

- Simulated samples of data are provided.
- The task is to classify into signal and background.

Signal $\tau\tau$ decay of a Higgs boson.

Background W, Z decays, $t\bar{t}$ products.

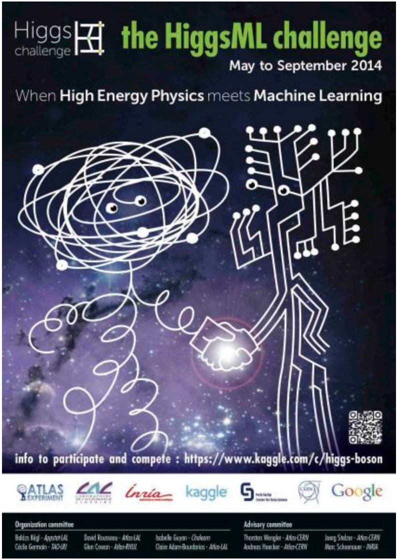
The Higgs Boson Machine Learning Challenge



The goal of the challenge is to explore the use of Machine Learning tools to improve the discovery significance of the experiment.

- Simulated samples of data are provided.
- The task is to classify into signal and background.

Signal	$\tau\tau$ decay of a Higgs boson.
Background	W, Z decays, $t\bar{t}$ products.



The goal of the challenge is to explore the use of Machine Learning tools to improve the discovery significance of the experiment.

- Simulated samples of data are provided.
- The task is to classify into signal and background.

Signal $\tau\tau$ decay of a Higgs boson.

Background W, Z decays, $t\bar{t}$ products.

Contents

Introduction

Machine Learning is a field concerned with developing algorithms that can *learn* from data.

Supervised learning Given a set of datapoints where the desired output is *known*, predict the output for unseen datapoints (classification, regression): *Generalize*.

Unsupervised learning Given a dataset where the output is *unknown*, discover structure in the data (clustering).

Introduction

Machine Learning is a field concerned with developing algorithms that can *learn* from data.

Supervised learning Given a set of datapoints where the desired output is *known*, predict the output for unseen datapoints (classification, regression): *Generalize*.

Unsupervised learning Given a dataset where the output is *unknown*, discover structure in the data (clustering).

Introduction

Machine Learning is a field concerned with developing algorithms that can *learn* from data.

Supervised learning Given a set of datapoints where the desired output is *known*, predict the output for unseen datapoints (classification, regression): *Generalize*.

Unsupervised learning Given a dataset where the output is *unknown*, discover structure in the data (clustering).

Introduction

Machine Learning is a field concerned with developing algorithms that can *learn* from data.

Supervised learning Given a set of datapoints where the desired output is *known*, predict the output for unseen datapoints (classification, regression): *Generalize*.

Unsupervised learning Given a dataset where the output is *unknown*, discover structure in the data (clustering).

Example: Regression problem

Given the set of data points $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}$, find a function $f(x) \to \mathbb{R}$ that *generalizes* them (ie is able to make a *good* prediction for new points generated by the same underlying model).

Good Maximize some score function.

- Typically split the known points into a *training set* and a *test set*.

Classification The same, but $f(x) \to \{-1, 1\}$

Example: Regression problem

Given the set of data points $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}$, find a function $f(x) \to \mathbb{R}$ that *generalizes* them (ie is able to make a *good* prediction for new points generated by the same underlying model).

Good Maximize some score function.

- Typically split the known points into a *training set* and a *test set*.

Classification The same, but $f(x) \to \{-1, 1\}$

Example: Regression problem

Given the set of data points $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}$, find a function $f(x) \to \mathbb{R}$ that *generalizes* them (ie is able to make a *good* prediction for new points generated by the same underlying model).

Good Maximize some score function.

- Typically split the known points into a *training set* and a *test set*.

Classification The same, but $f(x) \to \{-1, 1\}$

Example: Regression problem

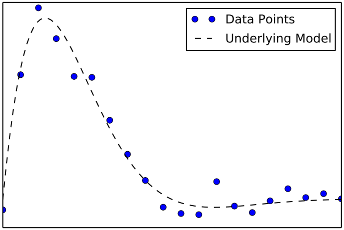
Given the set of data points $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}$, find a function $f(x) \to \mathbb{R}$ that *generalizes* them (ie is able to make a *good* prediction for new points generated by the same underlying model).

Good Maximize some score function.

- Typically split the known points into a *training set* and a *test set*.

Classification The same, but $f(x) \to \{-1, 1\}$

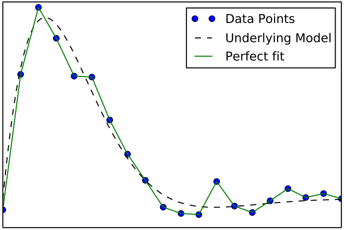
Example: Regression problem



Regression problem:

- Predicting the known data is not enough: We need *a priori* knowledge.
- Too strong assumptions about the model will lead to bad predictions.
- Different ML algorithms propose different solutions to finding an appropriate *bias/variance* compromise.

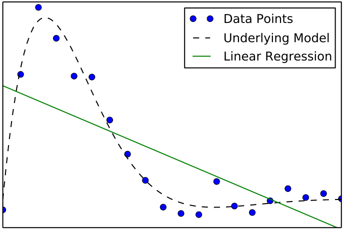
Example: Regression problem



Regression problem:

- Predicting the known data is not enough: We need *a priori* knowledge.
- Too strong assumptions about the model will lead to bad predictions.
- Different ML algorithms propose different solutions to finding an appropriate *bias/variance* compromise.

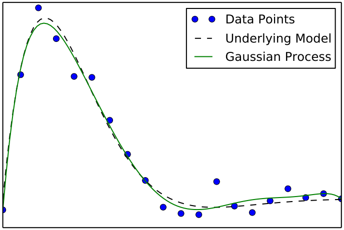
Example: Regression problem



Regression problem:

- Predicting the known data is not enough: We need *a priori* knowledge.
- Too strong assumptions about the model will lead to bad predictions.
- Different ML algorithms propose different solutions to finding an appropriate *bias/variance* compromise.

Example: Regression problem



Regression problem:

- Predicting the known data is not enough: We need *a priori* knowledge.
- Too strong assumptions about the model will lead to bad predictions.
- Different ML algorithms propose different solutions to finding an appropriate *bias/variance* compromise.

Dimensionality

One of the main challenges of the ML algorithms is to scale with the dimensionality of the feature space.

- Number of points required to sample the space $\sim 10^d$.
- Computational efficiency of some algorithms scales badly with d.
- Feature extraction or feature reduction are frequently used.

Dimensionality

One of the main challenges of the ML algorithms is to scale with the dimensionality of the feature space.

- Number of points required to sample the space $\sim 10^d$.
- Computational efficiency of some algorithms scales badly with d.
- Feature extraction or feature reduction are frequently used.

Dimensionality

One of the main challenges of the ML algorithms is to scale with the dimensionality of the feature space.

- Number of points required to sample the space $\sim 10^d$.
- Computational efficiency of some algorithms scales badly with d.
- Feature extraction or feature reduction are frequently used.

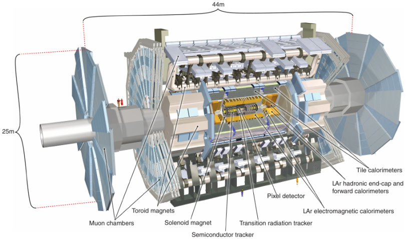
Dimensionality

One of the main challenges of the ML algorithms is to scale with the dimensionality of the feature space.

- Number of points required to sample the space $\sim 10^d$.
- Computational efficiency of some algorithms scales badly with d.
- Feature extraction or feature reduction are frequently used.

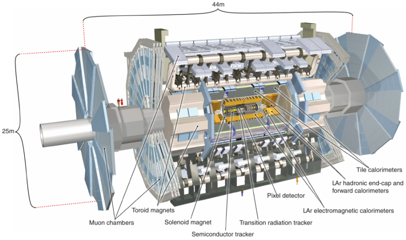
Contents

The detectors of ATLAS



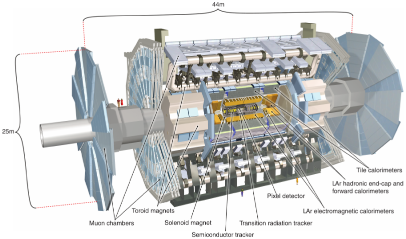
- Inner tracking detector: Momenta of charged particles.
- Electromagnetic calorimeter: Energy of photons and electrons.
- LAr calorimeter: Energy of hadrons and absorber.
- Muon spectrometer: Detects energy/momentum of muons.

The detectors of ATLAS



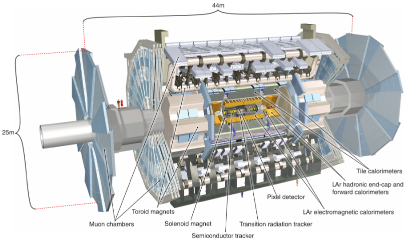
- Inner tracking detector: Momenta of charged particles.
- Electromagnetic calorimeter: Energy of photons and electrons.
- LAr calorimeter: Energy of hadrons and absorber.
- Muon spectrometer: Detects energy/momentum of muons.

The detectors of ATLAS



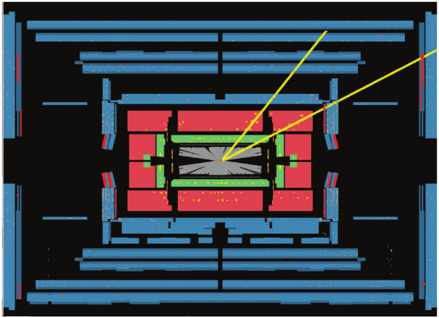
- Inner tracking detector: Momenta of charged particles.
- Electromagnetic calorimeter: Energy of photons and electrons.
- LAr calorimeter: Energy of hadrons and absorber.
- Muon spectrometer: Detects energy/momentum of muons.

The detectors of ATLAS



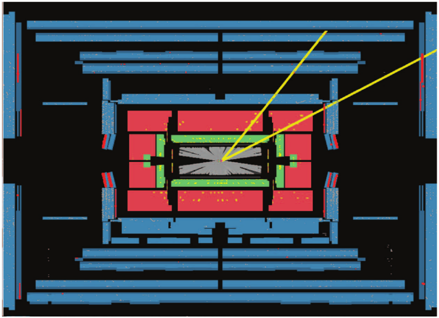
- Inner tracking detector: Momenta of charged particles.
- Electromagnetic calorimeter: Energy of photons and electrons.
- LAr calorimeter: Energy of hadrons and absorber.
- Muon spectrometer: Detects energy/momentum of muons.

The detectors of ATLAS



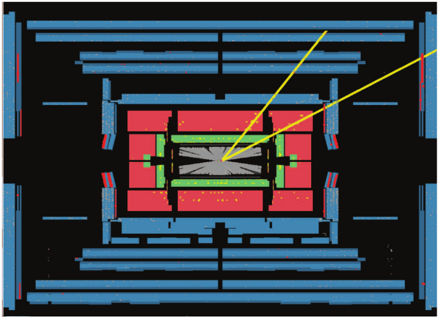
- Particle energy and momentum over 98% of the solid angle.
- Identification of electrons, muons and photons.
- Transverse momentum balance: neutrinos inferred.

The detectors of ATLAS



- Particle energy and momentum over 98% of the solid angle.
- Identification of electrons, muons and photons.
- Transverse momentum balance: neutrinos inferred.

The detectors of ATLAS



- Particle energy and momentum over 98% of the solid angle.
- Identification of electrons, muons and photons.
- Transverse momentum balance: neutrinos inferred.

Data simulation

The data provided for the challenge is simulated:

1. Random proton-proton collisions are simulated using Monte Carlo Event generators.
2. Cross sections are estimated from QCD theoretical calculations.
3. The resulting particles are tracked trough a virtual model of the detectors.
4. 30 physical observables are provided for the challenge.

Data simulation

The data provided for the challenge is simulated:

1. Random proton-proton collisions are simulated using Monte Carlo Event generators.
2. Cross sections are estimated from QCD theoretical calculations.
3. The resulting particles are tracked trough a virtual model of the detectors.
4. 30 physical observables are provided for the challenge.

Data simulation

The data provided for the challenge is simulated:

1. Random proton-proton collisions are simulated using Monte Carlo Event generators.
2. Cross sections are estimated from QCD theoretical calculations.
3. The resulting particles are tracked trough a virtual model of the detectors.
4. 30 physical observables are provided for the challenge.

Data simulation

The data provided for the challenge is simulated:

1. Random proton-proton collisions are simulated using Monte Carlo Event generators.
2. Cross sections are estimated from QCD theoretical calculations.
3. The resulting particles are tracked trough a virtual model of the detectors.
4. 30 physical observables are provided for the challenge.

Higgs fermionic decay

In the challenge, we are interested in the process (*signal*):

$$H \longrightarrow \tau^-\tau^+ \longrightarrow (l + 2\nu) + (\text{hadrons} + \nu)$$

$$l \in \{e^\pm, \mu^\pm), \nu \in \{\nu_e, \nu_\mu, \nu_\tau, \overline{\nu_e}, \overline{\nu_\mu}, \overline{\nu_\tau}\}$$

There are *background* events from:

$$Z \longrightarrow \tau^-\tau^+$$

$$t\bar{t} \rightarrow b\bar{b} + W^+W^- \rightarrow \tau^+\tau^- + \text{hadrons} + 2\nu$$

$$W \rightarrow (\text{wrong identification})$$

Higgs fermionic decay

In the challenge, we are interested in the process (*signal*):

$$H \longrightarrow \tau^-\tau^+ \longrightarrow (l + 2\nu) + (\text{hadrons} + \nu)$$

$$l \in \{e^\pm, \mu^\pm), \nu \in \{\nu_e, \nu_\mu, \nu_\tau, \overline{\nu_e}, \overline{\nu_\mu}, \overline{\nu_\tau}\}$$

There are *background* events from:

$$Z \longrightarrow \tau^-\tau^+$$

$$t\bar{t} \to b\bar{b} + W^+W^- \to \tau^+\tau^- + \text{hadrons} + 2\nu$$

$$W \to (\text{wrong identification})$$

Contents

The classification problem

Let $\mathcal{D}$ be the training sample:

$$\mathcal{D} = \{(x_1, y_1, w_1), ..., (x_n, y_n, w_n)\}$$

where:

- $x_i \in \mathbb{R}^d$: Feature vector.
- $y_i \in \{b = \text{"}background\text{"}, s = \text{"}signal\text{"}\}$: Label
- $w_i \in \mathbb{R}^+$: Weight.

Find a classifier $g : \mathbb{R}^d \to \{b, s\}$ that maximizes the *Approximate Median Significance*.

The classification problem

Let $\mathcal{D}$ be the training sample:

$$\mathcal{D} = \{(x_1, y_1, w_1), ..., (x_n, y_n, w_n)\}$$

where:

- $x_i \in \mathbb{R}^d$: Feature vector.
- $y_i \in \{b \equiv "background", s \equiv "signal"\}$: Label
- $w_i \in \mathbb{R}^+$: Weight.

Find a classifier $g : \mathbb{R}^d \to \{b, s\}$ that maximizes the *Approximate Median Significance*.

The classification problem

Let $\mathcal{D}$ be the training sample:

$$\mathcal{D} = \{(x_1, y_1, w_1), ..., (x_n, y_n, w_n)\}$$

where:

- $x_i \in \mathbb{R}^d$: Feature vector.
- $y_i \in \{b \equiv \text{"}background\text{"}, s \equiv \text{"}signal\text{"}\}$: Label
- $w_i \in \mathbb{R}^+$: Weight.

Find a classifier $g : \mathbb{R}^d \to \{b, s\}$ that maximizes the *Approximate Median Significance*.

Chosen method

The first submission uses a Gradient Boosting Classifier:

- Use many weak algorithms to produce a strong prediction.
- Similar to the ATLAS analysis strategy.
- Train different classifiers for different numbers of jets and for the case of mass not provided.

Chosen method

The first submission uses a Gradient Boosting Classifier:

- Use many weak algorithms to produce a strong prediction.
- Similar to the ATLAS analysis strategy.
- Train different classifiers for different numbers of jets and for the case of mass not provided.

Chosen method

The first submission uses a Gradient Boosting Classifier:

- Use many weak algorithms to produce a strong prediction.
- Similar to the ATLAS analysis strategy.
- Train different classifiers for different numbers of jets and for the case of mass not provided.

Gradient Boosting Classifier Formulation

$$g(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

- h_m weak learners (decision trees).
- h_m chosen to minimize some loss function $L(y_i, x_i)$ at each iteration:

$$g_m(x) = g_{m-1}(x) + \arg\min_h \sum_{i=1}^{n} L(y_i, g_{m-1}(x_i) - h(x))$$

- The minimization is performed by the steepest descent method:

$$g_m(x) = g_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_g L(y_i, g_{m-1}(x_i))$$

$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L\left(y_i, g_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right)$$

Gradient Boosting Classifier Formulation

$$g(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

- h_m weak learners (decision trees).
- h_m chosen to minimize some loss function $L(y_i, x_i)$ at each iteration:

$$g_m(x) = g_{m-1}(x) + \arg\min_{h} \sum_{i=1}^{n} L(y_i, g_{m-1}(x_i) - h(x))$$

- The minimization is performed by the steepest descent method:

$$g_m(x) = g_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_g L(y_i, g_{m-1}(x_i))$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, g_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right)$$

Gradient Boosting Classifier Formulation

$$g(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

- h_m weak learners (decision trees).
- h_m chosen to minimize some loss function $L(y_i, x_i)$ at each iteration:

$$g_m(x) = g_{m-1}(x) + \arg\min_h \sum_{i=1}^{n} L(y_i, g_{m-1}(x_i) - h(x))$$

- The minimization is performed by the steepest descent method:

$$g_m(x) = g_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_g L(y_i, g_{m-1}(x_i))$$

$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L\left(y_i, g_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right)$$

Results

The first submission obtained an AMS score of 3.376.

	Score
Random Submission	0.58
Simple Window	1.54
Naive Bayes	2.06
Simple Boosted Trees	3.25
AdaBoost	3.34
My submission	3.38
Best submission	3.81
ATLAS (real significance)	4.1

Next steps

- Reduce overfitting.
- Tweak parameters.
- Explore other methods.

Next steps

- Reduce overfitting.
- Tweak parameters.
- Explore other methods.

Next steps

- Reduce overfitting.
- Tweak parameters.
- Explore other methods.

Sources I

- Cécile Germain Isabelle Guyon Balázs Kégl David Rousseau Claire Adam-Bourdarios, Glen Cowan.
  Learning to discover: the Higgs boson machine learning challenge, 2014.
  URL http://higgsml.lal.in2p3.fr/documentation/.
- ATLAS collaboration et al.
  Evidence for Higgs boson decays to the $\tau^{+}\ \tau^{-}$ final state with the ATLAS detector.
  ATLAS-CONF-2013-108, 2013.
- Georges Aad, T Abajyan, B Abbott, J Abdallah, S Abdel Khalek, AA Abdelalim, O Abdinov, R Aben, B Abi, M Abolins, et al.
  A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider.
  Science, 338(6114):1576–1582, 2012.

Sources II

The CMS Collaboration.
Evidence for the direct decay of the 125 GeV Higgs boson to fermions.
Nat Phys, advance online publication, Jun 2014.
ISSN 1745-2481.
URL http://dx.doi.org/10.1038/nphys3005.
Letter.

P K Sinervo.
Signal Significance in Particle Physics.
(hep-ex/0208005. CDF-PUB-STATISTICS-PUBLIC-6031), Aug 2002.

P F Harrison.
Blind analysis.
Journal of Physics G: Nuclear and Particle Physics, 28(10):2679, 2002.
URL http://stacks.iop.org/0954-3899/28/i=10/a=312.

Sources III

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al.
Scikit-learn: Machine learning in Python.
The Journal of Machine Learning Research, 12:2825–2830, 2011.