Classication of data from the ATLAS experiment

Michael Janschek

HHU Düsseldorf

22. März 2016

Contents

1 Introduction
- ATLAS
- Data processing of ATLAS

2 The Task
- The Higgs Boson Machine Learning Challenge
- Approximate Median Significance

3 Methods of Classification
- Logistic Regression
- K Nearest Neighbor
- XGBoost

4 Conclusion
- Results

Introduction

A Toroidal LHC ApparatuS (ATLAS)

- Registers ∼40 million particle collisions (called events) per second.
- Aims to investigate four major topics in physics.
- Discovered the Higgs Boson in 2012.
- Tests predicted properties of the Higgs boson.

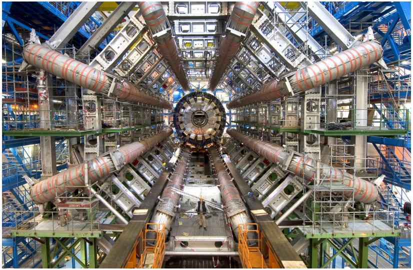

Figure: The ATLAS detector. [ATL16]

Data processing of ATLAS

- Three levels of filtering, reducing events from 40 million to 200 per second. (The *trigger*)
- Regions in feature space are picked for further analysis. They are called the *selection region*.
- Recorded and compressed events that are part of this region are reconstructed for the current task.

[ATL16]

The Task

The Higgs Boson Machine Learning Challenge

- Perform event selection on simulated data.
- Classify events correctly to maximize evaluation metric.
- Searched *signal events* are $\mathrm{H} \rightarrow \tau\tau$ decay, *background events* originate from three other processes.

The data

Simulated data allows precise labeling and importance weighting of events.

250000 events for training, 550000 for testing.

- 1st feature is event ID
- 30 features with physics information
- 2 features with weighting and labeling (exclusive for training data)

Challenge data can be recreated from `opendata.cern.ch`

Approximate Median Significance

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

where

- s is the sum of weights of *true* signals
- b is the sum of weights of *false* signals
- $b_{reg} = 10$ is an artificial shift to b

Related to statistical significance.
Used in particle physics to optimize the selection region.

Methods of Classification

Logistic Regression

- Builds score out of weighted features.
- Uses score in logistic function.
- Predicts the *probability* of test data being part of a class.

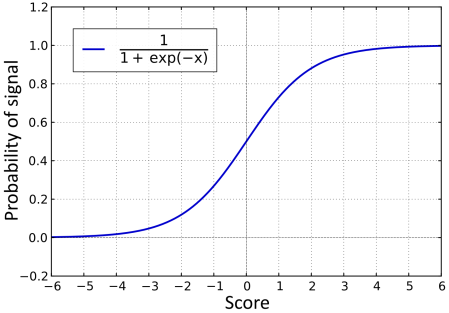

Figure: The standard logistic function [log]

K Nearest Neighbor

Compare new data point directly to *nearest* training data points.

K Nearest Neighbor

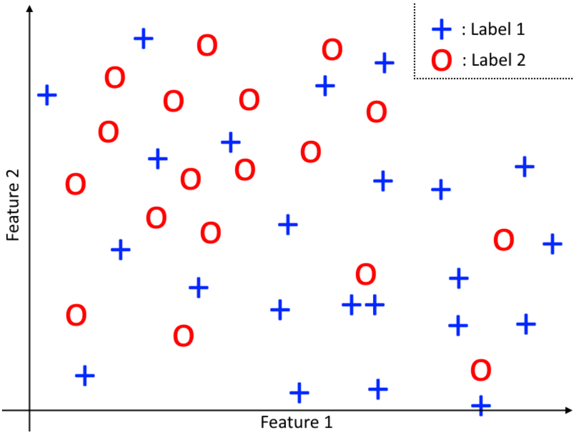

Figure: Initializing kNN

K Nearest Neighbor

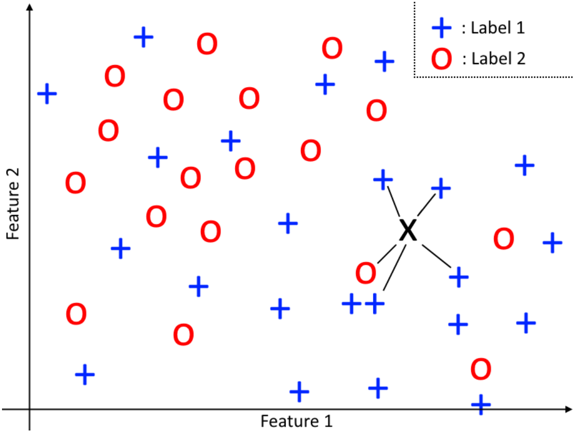

Figure: Prediction in kNN (1)

K Nearest Neighbor

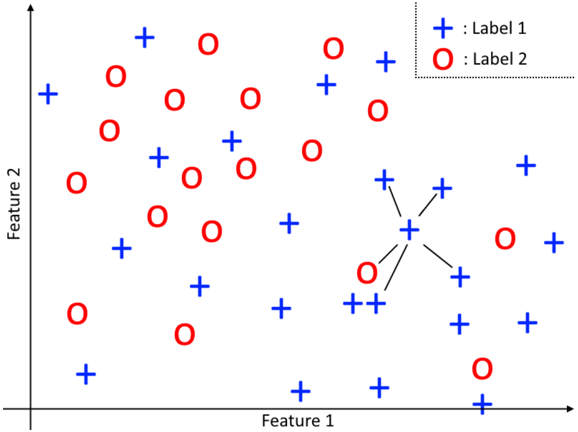

Figure: Prediction in kNN (2)

XGBoost

- Gradient Boosting classifier developed by Tianqi Chen.
- Shared early with other participants and grew popular.
- Was acknowledged with the *HEP meets ML Award*.
  - Prediction performance
  - Documentation
  - CPU and memory demands
  - simplicity/straightforwardness of approach

Boosting

$$
\begin{aligned}
f_1(x) &= y - error \\
error &= y - f_1(x) \\
f_2(x) &= y - f_1(x) \\
f_1(x) + f_2(x) &= y - error_2 \\
&\dots
\end{aligned}
$$

We combine an ensemble of *weak learners* $f_i(x)$ into one *strong learner* $F(x)$.

Gradient Boosting

$$\text{square loss: } l = \sum_i^n \frac{(f_i(x)-y)^2}{2}$$

$$\frac{\partial l}{\partial f_1(x)} = f_1(x) - y$$

$$-\frac{\partial l}{\partial f_1(x)} = y - f_1(x)$$

$$-\frac{\partial l}{\partial f_1(x)} = f_2(x)$$

Residuals can be interpreted as *negative gradients*.

If we fit *weak learners* to *negative gradients*, it allows to minimize the loss function using *gradient descent*.

One can show that a gradient boosting algorithm can be constructed for any loss function [Li].

XGBoost

Regularized objective function

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

- l is any loss function.
- $\Omega(f_k)$ measures complexity of classifier f_k.
- As L is minimized, so are l and Ω.

XGBoost

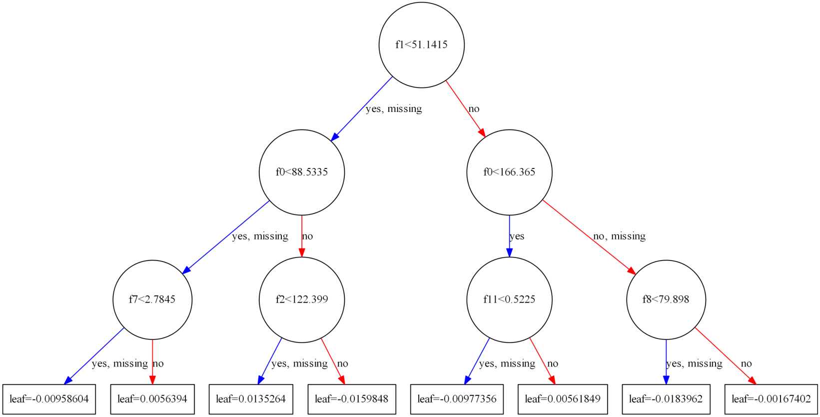

Figure: A decision tree generated by XGBoost.

Conclusion

Observations after the challenge:

- Linear models performed badly in the challenge.
- K Nearest Neighbors nearly beats most of the challenges benchmarks.
- Boosting methods and neural networks performed well.

Results

Classifier	AMS	rank
Logistic Regression	2.06934	1429
k Nearest Neighbor	3.18323	996
XGBoost	3.71268	65
XGBoost original	3.71885	45
Winning submission	3.80581	1

Table: Performances of used methods

1785 participating teams in total.

Results

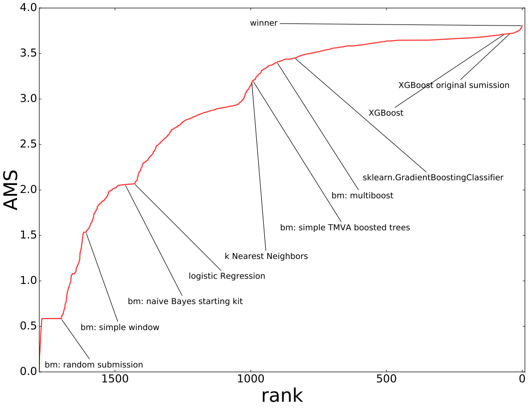

Figure: AMS of all final submissions in the challenge. [Hig]

References I

[ATL16] Official homepage of atlas.
URL http://www.atlas.ch, 2016.
Accessed: 2016-02-15.

[Hig] Higgs boson machine learning challenge.
https://www.kaggle.com/c/higgs-boson.
Accessed: 2016-01-03.

[Li] Cheng Li.
A gentle introduction to gradient boosting.

[log] Wikipedia.
URL https://de.wikipedia.org/wiki/Logistische_Regression.
Accessed: 2016-03-20.