INSTITUT FÜR INFORMATIK
Computer Vision, Computer Graphics
and Pattern Recognition

Universitätsstr. 1 D–40225 Düsseldorf

Classification of data from the ATLAS experiments

Michael Janschek

Bachelor Thesis

Beginn der Arbeit:	10. Dezember 2015
Abgabe der Arbeit:	10. März 2016
Gutachter:	Prof. Dr. Stefan Harmeling
	Prof. Dr. Stefan Conrad

Erklärung

Hiermit versichere ich, dass ich diese Bachelor Thesis selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 10. März 2016

Michael Janschek

Abstract

Hier kommt eine ca. einseitige Zusammenfassung der Arbeit rein.

Contents

1 Introduction

1.1 Motivation

1.2 Goal

1.3 Overview

blablalba

2 The Higgs Boson Machine Learning Challenge

2.1 The Data

2.2 The formal problem

2.3 The Evaluation

3 Methods of classification

3.1 Logistic regression

4 Using logistic Regression on Toydata to get a high AMS

```
{\color{incolor}In [{\color{incolor}1}]:} \PY{k+kn}{import} \PY{n+nn}{nump
        \PY{k+kn}{import} \PY{n+nn}{matplotlib}\PY{n+nn}{.}\PY{n+nn}{pyplo
        \PY{k+kn}{import} \PY{n+nn}{math}
        \PY{k+kn}{from} \PY{n+nn}{sklearn} \PY{k}{import} \PY{n}{linear\PY
```

Data shall have the form of $[w, y, x_1, x_2]$ where

- w is a weight in the intervall $[0, 1)$
- y is the label "0" for "background" or "1" for "signal"
- x_n are randomly generated features with respect to the label

```
{\color{incolor}In [{\color{incolor}2}]:} \PY{k}{def} \PY{n+nf}{generateFe
            \PY{k}{if} \PY{n}{label} \PY{o+ow}{is} \PY{l+m+mi}{1}\PY{p}{:}
                \PY{n}{mu} \PY{o}{=} \PY{n}{mu\PYZus{}s}
                \PY{n}{sigma} \PY{o}{=} \PY{n}{sigma\PYZus{}s}
            \PY{k}{else}\PY{p}{:}
                \PY{n}{mu} \PY{o}{=} \PY{n}{mu\PYZus{}b}
                \PY{n}{sigma} \PY{o}{=} \PY{n}{sigma\PYZus{}b}
            \PY{k}{return} \PY{n}{np}\PY{o}{.}\PY{n}{random}\PY{o}{.}\PY{r
```

Approximate Median Significance (AMS) defined as:

$$AMS = \sqrt{2(s + b + b_r)log[1 + (s/(b + b_{reg}))] - s}$$

where $b_{reg} = 10$ is a regulization term (set by the contest),
$b = \sum_{i=1}^{n} w_i, y_i = 0$ is sum of weighted background (incorrectly classified as signal),
$s = \sum_{i=1}^{n} w_i, y_i = 1$ is sum of weighted signals (correctly classified as signal),
log is natural logarithm

```
{\color{incolor}In [{\color{incolor}3}]:} \PY{k}{def} \PY{n+nf}{calcAMS}\P

            \PY{n}{br} \PY{o}{=} \PY{l+m+mf}{10.0}
            \PY{n}{radicand} \PY{o}{=} \PY{l+m+mi}{2} \PY{o}{*}\PY{p}{(} \
            \PY{k}{if} \PY{n}{radicand} \PY{o}{\PYZlt{}} \PY{l+m+mi}{0}\PY
                \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZsq{}}\PY{l+s}{radica
                \PY{n}{exit}\PY{p}{(}\PY{p}{)}
            \PY{k}{else}\PY{p}{:}
                \PY{k}{return} \PY{n}{math}\PY{o}{.}\PY{n}{sqrt}\PY{p}{(}\
```

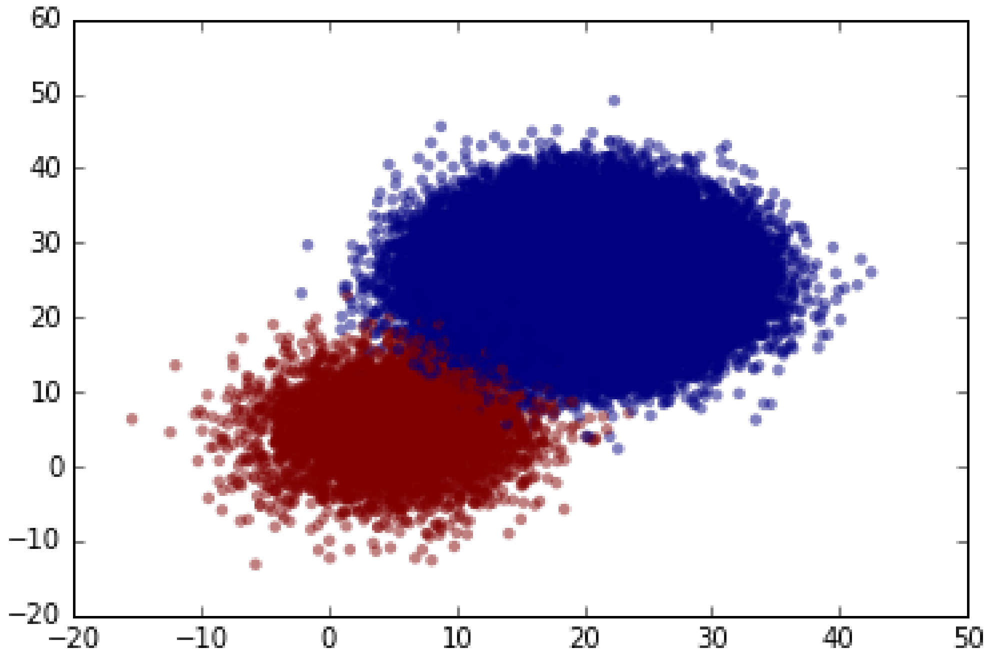
```
{\color{incolor}In [{\color{incolor}4}]:} \PY{k}{def} \PY{n+nf}{calcWeight
            \PY{n}{s} \PY{o}{=} \PY{l+m+mi}{0}
            \PY{n}{b} \PY{o}{=} \PY{l+m+mi}{0}
            \PY{k}{for} \PY{n}{j} \PY{o+ow}{in} \PY{n+nb}{list}\PY{p}{(}\P
                \PY{n}{pred} \PY{o}{=} \PY{n}{preds}\PY{p}{[}\PY{n}{j}\PY{
                \PY{n}{label} \PY{o}{=} \PY{n}{labels}\PY{p}{[}\PY{n}{j}\P
                \PY{n}{weight} \PY{o}{=} \PY{n}{weights}\PY{p}{[}\PY{n}{j}
                \PY{k}{if} \PY{n}{pred} \PY{o}{\PYZgt{}} \PY{l+m+mf}{0.}\P
                    \PY{k}{if} \PY{n}{label} \PY{o}{\PYZgt{}} \PY{l+m+mf}{
                        \PY{n}{s} \PY{o}{+}\PY{o}{=} \PY{n}{weight}
                    \PY{k}{else}\PY{p}{:}
                        \PY{n}{b} \PY{o}{+}\PY{o}{=} \PY{n}{weight}
            \PY{k}{return} \PY{n}{s}\PY{p}{,}\PY{n}{b}
```

actually generate data

```
{\color{incolor}In [{\color{incolor}5}]:} \PY{n}{n} \PY{o}{=} \PY{l+m+mi}{
        \PY{n}{s\PYZus{}prob} \PY{o}{=} \PY{l+m+mf}{0.05} \PY{c}{\PYZsh{}p
        \PY{n}{weights} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{random}\PY{o}{
        \PY{n}{labels} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{zeros}\PY{p}{(}
        \PY{n}{x\PYZus{}1} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{zeros}\PY{p
        \PY{n}{x\PYZus{}2} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{zeros}\PY{p

        \PY{k}{for} \PY{n}{i} \PY{o+ow}{in} \PY{n+nb}{range}\PY{p}{(}\PY{l
            \PY{k}{if} \PY{n}{weights}\PY{p}{[}\PY{n}{i}\PY{p}{]} \PY{o}{\
                \PY{n}{label} \PY{o}{=} \PY{l+m+mi}{1}
            \PY{k}{else}\PY{p}{:}
                \PY{n}{label} \PY{o}{=} \PY{l+m+mi}{0}
            \PY{n}{labels}\PY{p}{[}\PY{n}{i}\PY{p}{]} \PY{o}{=} \PY{n}{lab
            \PY{n}{x\PYZus{}1}\PY{p}{[}\PY{n}{i}\PY{p}{]}\PY{o}{=}\PY{n}{g
            \PY{n}{x\PYZus{}2}\PY{p}{[}\PY{n}{i}\PY{p}{]}\PY{o}{=}\PY{n}{g
```

visualize

```
{\color{incolor}In [{\color{incolor}29}]:} \PY{o}{\PYZpc{}}\PY{k}{pylab} i
         \PY{n}{plt}\PY{o}{.}\PY{n}{scatter}\PY{p}{(}\PY{n}{x\PYZus{}1}\PY

Populating the interactive namespace from numpy and matplotlib
    \end{Verbatim}

            \begin{Verbatim}[commandchars=\\\{\}]
{\color{outcolor}Out[{\color{outcolor}29}]:} <matplotlib.collections.PathC
```




```
{\color{incolor}In [{\color{incolor}7}]:} \PY{k}{def} \PY{n+nf}{splitList}
            \PY{n}{aList} \PY{o}{=} \PY{n}{xList}\PY{p}{[}\PY{p}{:}\PY{n}{
            \PY{n}{bList} \PY{o}{=} \PY{n}{xList}\PY{p}{[}\PY{n}{n}\PY{p}{
            \PY{k}{return} \PY{n}{aList}\PY{p}{,}\PY{n}{bList}
```

split toydata into training- and testset for the classifier

```
{\color{incolor}In [{\color{incolor}8}]:} \PY{n}{n\PYZus{}train} \PY{o}{=}

        \PY{n}{train\PYZus{}x\PYZus{}1}\PY{p}{,}\PY{n}{test\PYZus{}x\PYZus
        \PY{n}{train\PYZus{}x\PYZus{}2}\PY{p}{,}\PY{n}{test\PYZus{}x\PYZus
        \PY{n}{train\PYZus{}labels}\PY{p}{,}\PY{n}{test\PYZus{}labels} \PY
        \PY{n}{test\PYZus{}weights} \PY{o}{=} \PY{n}{splitList}\PY{p}{(}\P
```

For Comparison, we calculate the best possible AMS (case: every signal correctly detected)

```
{\color{incolor}In [{\color{incolor}9}]:} \PY{k}{def} \PY{n+nf}{calcMaxAMS
            \PY{n}{s}\PY{p}{,}\PY{n}{b} \PY{o}{=} \PY{n}{calcWeightSums}\P
            \PY{n}{ams} \PY{o}{=} \PY{n}{calcAMS}\PY{p}{(}\PY{n}{s}\PY{p}{
            \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{Maximum AM
            \PY{k}{return} \PY{n}{ams}
```

```
{\color{incolor}In [{\color{incolor}10}]:} \PY{n}{calcMaxAMS}\PY{p}{(}\PY{

Maximum AMS possible with this Data: 20.002095901254826
    \end{Verbatim}

            \begin{Verbatim}[commandchars=\\\{\}]
{\color{outcolor}Out[{\color{outcolor}10}]:} 20.002095901254826
```

we initialize the Logistic Regression Classifier, shape the input-data and fit the model

```
{\color{incolor}In [{\color{incolor}11}]:} \PY{n}{logReg} \PY{o}{=} \PY{n}

         \PY{n}{train\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{train\PYZus{}labels} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{a
         \PY{n}{test\PYZus{}labels} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{ar

         \PY{n}{logReg}\PY{o}{.}\PY{n}{fit}\PY{p}{(}\PY{n}{train\PYZus{}x}

         \PY{n}{logReg}\PY{o}{.}\PY{n}{sparsify}\PY{p}{(}\PY{p}{)}

         \PY{n}{predProb} \PY{o}{=} \PY{n}{logReg}\PY{o}{.}\PY{n}{predict\
         \PY{n}{pred} \PY{o}{=} \PY{n}{logReg}\PY{o}{.}\PY{n}{predict}\PY{
         \PY{n}{score} \PY{o}{=} \PY{n}{logReg}\PY{o}{.}\PY{n}{score}\PY{p

         \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{Score:}\PY{l+

Score: 0.997377777778
    \end{Verbatim}

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}12}]:} \PY{n}{s}\PY{p}{,}\PY{n}{b} \PY
         \PY{n}{calcAMS}\PY{p}{(}\PY{n}{s}\PY{p}{,}\PY{n}{b}\PY{p}{)}

{\color{outcolor}Out[{\color{outcolor}12}]:} 10.596863251598934
```

We successfully tested logistic Regression, now let's use it on actual CERN-Data.

```
{\color{incolor}In [{\color{incolor}13}]:} \PY{k+kn}{import} \PY{n+nn}{Kag

{\color{incolor}In [{\color{incolor}14}]:} \PY{n}{csvDict}\PY{p}{,}\PY{n}{

Reading csv file  /home/garg/Dokumente/workspace/BA\_git/Data/Atlas-higgs-
    \end{Verbatim}
```

```
    Trainingset has key ``t''\\
Public Testset has key ``p'' (note: ``p'' won't work, using private
Testset (``v''))

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}30}]:} \PY{k}{def} \PY{n+nf}{getFeatur
             \PY{n}{trainFeature} \PY{o}{=} \PY{n}{KaggleData}\PY{o}{.}\PY
             \PY{n}{testFeature} \PY{o}{=} \PY{n}{KaggleData}\PY{o}{.}\PY{
             \PY{k}{return} \PY{n}{trainFeature}\PY{p}{,} \PY{n}{testFeatu

{\color{incolor}In [{\color{incolor}31}]:} \PY{n}{train\PYZus{}eventList}\
         \PY{n}{train\PYZus{}labels}\PY{p}{,}\PY{n}{test\PYZus{}labels} \P
         \PY{n}{test\PYZus{}weights} \PY{o}{=} \PY{n}{getFeatureSets}\PY{p
```

We observe the relation Label <=> Weight

```
{\color{incolor}In [{\color{incolor}32}]:} \PY{n}{signal\PYZus{}sum} \PY{o
         \PY{n}{background\PYZus{}sum} \PY{o}{=} \PY{n+nb}{int}\PY{p}{(}\P
         \PY{n}{signal\PYZus{}weight} \PY{o}{=} \PY{l+m+mi}{0}
         \PY{n}{background\PYZus{}weight} \PY{o}{=} \PY{l+m+mi}{0}
         \PY{k}{for} \PY{n}{i} \PY{o+ow}{in} \PY{n+nb}{range}\PY{p}{(}\PY{
             \PY{k}{if} \PY{n}{test\PYZus{}labels}\PY{p}{[}\PY{n}{i}\PY{p}
                 \PY{n}{signal\PYZus{}weight} \PY{o}{+}\PY{o}{=} \PY{n}{te
             \PY{k}{else}\PY{p}{:}
                 \PY{n}{background\PYZus{}weight} \PY{o}{+}\PY{o}{=} \PY{n
         \PY{n+nb}{print}\PY{p}{(}\PY{n}{background\PYZus{}weight}\PY{o}{/
         \PY{n+nb}{print}\PY{p}{(}\PY{n}{signal\PYZus{}weight}\PY{o}{/}\PY

1.38702756616
0.00450270106462
    \end{Verbatim}

    We can observe, that False signals will be weighted a lot heavier than
True signals.

If a classifier achieved a higher AMS while detecting less signals, we
can make statements about the usabilty of the features, the classifier
used.

We choose features with beneficial properties for classifying.

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}33}]:} \PY{n}{train\PYZus{}DER\PYZus{}
         \PY{n}{train\PYZus{}DER\PYZus{}pt\PYZus{}ratio\PYZus{}lep\PYZus{}
```

Using DER_mass_MMC was not allowed in the former contest, we use it here anyway to test our classifier

```
{\color{incolor}In [{\color{incolor}34}]:} \PY{n}{train\PYZus{}DER\PYZus{}

{\color{incolor}In [{\color{incolor}35}]:} \PY{n}{train\PYZus{}labels} \PY
         \PY{n}{test\PYZus{}labels} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{ar

{\color{incolor}In [{\color{incolor}36}]:} \PY{n}{calcMaxAMS}\PY{p}{(}\PY{
         \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{True Signals:

Maximum AMS possible with this Data: 67.71112289514183
True Signals: 153683
    \end{Verbatim}

    We start with one feature and add more with every regression to see
improvement of the AMS

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}37}]:} \PY{k}{def} \PY{n+nf}{logisticR
             \PY{n}{logReg} \PY{o}{=} \PY{k}{None}
             \PY{n}{logReg} \PY{o}{=} \PY{n}{linMod}\PY{o}{.}\PY{n}{Logist

             \PY{n}{logReg}\PY{o}{.}\PY{n}{fit}\PY{p}{(}\PY{n}{train\PYZus

             \PY{n}{logReg}\PY{o}{.}\PY{n}{sparsify}\PY{p}{(}\PY{p}{)}

             \PY{n}{predProb} \PY{o}{=} \PY{n}{logReg}\PY{o}{.}\PY{n}{pred
             \PY{n}{pred} \PY{o}{=} \PY{n}{logReg}\PY{o}{.}\PY{n}{predict}
             \PY{n}{signals} \PY{o}{=} \PY{n+nb}{int}\PY{p}{(}\PY{n}{pred}

             \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{signals r
             \PY{k}{if} \PY{n}{signals} \PY{o+ow}{is} \PY{o+ow}{not} \PY{l
                 \PY{n}{s}\PY{p}{,}\PY{n}{b} \PY{o}{=} \PY{n}{calcWeightSu
                 \PY{n}{ams} \PY{o}{=} \PY{n}{calcAMS}\PY{p}{(}\PY{n}{s}\P
             \PY{k}{else}\PY{p}{:}
                 \PY{n}{ams} \PY{o}{=} \PY{l+m+mi}{0}
             \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{AMS:}\PY{
             \PY{k}{return} \PY{n}{predProb}\PY{p}{,}\PY{n}{pred}\PY{p}{,}

{\color{incolor}In [{\color{incolor}38}]:} \PY{n}{train\PYZus{}x} \PY{o}{=
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{pred} \PY{o}{=} \PY{n}{logisticReg}\PY{p}{(}\PY{n}{train\P
         \PY{n}{pred}\PY{o}{.}\PY{n}{cumsum}\PY{p}{(}\PY{p}{)}

signals read: 105546
AMS: 1.3069521546253222
    \end{Verbatim}
```

```
            \begin{Verbatim}[commandchars=\\\{\}]
{\color{outcolor}Out[{\color{outcolor}38}]:} array([      0.,       0.,

{\color{incolor}In [{\color{incolor}39}]:} \PY{k}{def} \PY{n+nf}{logRegFor
             \PY{k}{for} \PY{n}{feature} \PY{o+ow}{in} \PY{n}{fList}\PY{p}
                 \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{Featu
                 \PY{n}{trainList\PYZus{}x}\PY{p}{,}\PY{n}{testList\PYZus{
                 \PY{n}{train\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{r
                 \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}
                 \PY{n}{logisticReg}\PY{p}{(}\PY{n}{train\PYZus{}x}\PY{p}{

{\color{incolor}In [{\color{incolor}40}]:} \PY{l+s+sd}{\PYZdq{}\PYZdq{}\PY
         \PY{l+s+sd}{logRegFor([\PYZdq{}PRI\PYZus{}met\PYZdq{},}
         \PY{l+s+sd}{           \PYZdq{}PRI\PYZus{}met\PYZus{}sumet\PYZdq{
         \PY{l+s+sd}{           \PYZdq{}PRI\PYZus{}tau\PYZus{}pt\PYZdq{},}
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}met\PYZus{}phi\PYZus{}c
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}pt\PYZus{}ratio\PYZus{}
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}sum\PYZus{}pt\PYZdq{},}
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}pt\PYZus{}h\PYZdq{},}
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}mass\PYZus{}transverse\
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}mass\PYZus{}MMC\PYZdq{}
         \PY{l+s+sd}{           \PYZdq{}DER\PYZus{}mass\PYZus{}jet\PYZus{}
         \PY{l+s+sd}{           \PYZdq{}PRI\PYZus{}jet\PYZus{}all\PYZus{}p
         \PY{l+s+sd}{\PYZdq{}\PYZdq{}\PYZdq{}}

{\color{outcolor}Out[{\color{outcolor}40}]:} '\textbackslash{}nlogRegFor([

{\color{incolor}In [{\color{incolor}41}]:} \PY{n}{train\PYZus{}PRI\PYZus{}
         \PY{n}{train\PYZus{}DER\PYZus{}met\PYZus{}phi\PYZus{}centrality}\
         \PY{n}{train\PYZus{}DER\PYZus{}pt\PYZus{}h}\PY{p}{,}\PY{n}{test\P
         \PY{n}{train\PYZus{}DER\PYZus{}pt\PYZus{}ratio\PYZus{}lep\PYZus{}
         \PY{n}{train\PYZus{}DER\PYZus{}mass\PYZus{}transverse\PYZus{}met\
```

we are able to achieve a higher AMS by adjusting the decision-threshold (around 0.25)

```
{\color{incolor}In [{\color{incolor}42}]:} \PY{k}{def} \PY{n+nf}{bestThres
             \PY{n}{thresh} \PY{o}{=} \PY{l+m+mi}{0}
             \PY{n}{maxAMS} \PY{o}{=} \PY{l+m+mi}{0}
             \PY{n}{maxThresh} \PY{o}{=} \PY{l+m+mi}{0}
             \PY{k}{for} \PY{n}{thresh} \PY{o+ow}{in} \PY{n}{np}\PY{o}{.}\
                 \PY{n}{newPred} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{zeros
                 \PY{k}{for} \PY{n}{i} \PY{o+ow}{in} \PY{n+nb}{range}\PY{p
                     \PY{k}{if} \PY{n}{predProb}\PY{p}{[}\PY{n}{i}\PY{p}{]
                         \PY{n}{newPred}\PY{p}{[}\PY{n}{i}\PY{p}{]}\PY{o}{
                 \PY{n}{s}\PY{p}{,}\PY{n}{b} \PY{o}{=} \PY{n}{calcWeightSu
```

```
                  \PY{n}{ams} \PY{o}{=} \PY{n}{calcAMS}\PY{p}{(}\PY{n}{s}\P
                  \PY{k}{if} \PY{n}{ams} \PY{o}{\PYZgt{}} \PY{n}{maxAMS}\PY
                      \PY{n}{maxThresh} \PY{o}{=} \PY{n}{thresh}
                      \PY{n}{maxAMS} \PY{o}{=} \PY{n}{ams}
                      \PY{n}{signals} \PY{o}{=} \PY{n+nb}{int}\PY{p}{(}\PY{
              \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{Maximum A
              \PY{n+nb}{print}\PY{p}{(}\PY{l+s}{\PYZdq{}}\PY{l+s}{Signals r

{\color{incolor}In [{\color{incolor}43}]:} \PY{n}{train\PYZus{}x} \PY{o}{=
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{predProb}\PY{p}{,}\PY{n}{pred} \PY{o}{=} \PY{n}{logisticRe
         \PY{n}{bestThreshold}\PY{p}{(}\PY{n}{predProb}\PY{p}{)}

signals read: 88314
AMS: 1.319232266698236
Maximum AMS: 1.4163299443927306 with threshold 0.256565656566
Signals read: 266273
    \end{Verbatim}

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}44}]:} \PY{n}{train\PYZus{}x} \PY{o}{=
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{predProb}\PY{p}{,}\PY{n}{pred} \PY{o}{=} \PY{n}{logisticRe
         \PY{n}{bestThreshold}\PY{p}{(}\PY{n}{predProb}\PY{p}{)}

signals read: 70612
AMS: 1.1512119152387188
Maximum AMS: 1.3390609562029994 with threshold 0.224242424242
Signals read: 296710
    \end{Verbatim}

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}45}]:} \PY{n}{train\PYZus{}x} \PY{o}{=
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{predProb}\PY{p}{,}\PY{n}{pred} \PY{o}{=} \PY{n}{logisticRe
         \PY{n}{bestThreshold}\PY{p}{(}\PY{n}{predProb}\PY{p}{)}

signals read: 105546
AMS: 1.3069521546253222
Maximum AMS: 1.4048680561650573 with threshold 0.256565656566
Signals read: 274385
    \end{Verbatim}

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor}49}]:} \PY{n}{train\PYZus{}x} \PY{o}{=
         \PY{n}{test\PYZus{}x} \PY{o}{=} \PY{n}{np}\PY{o}{.}\PY{n}{array}\
         \PY{n}{predProb}\PY{p}{,}\PY{n}{pred} \PY{o}{=} \PY{n}{logisticRe
```

```
signals read: 70612
AMS: 1.1512119152387188
    \end{Verbatim}

    \begin{Verbatim}[commandchars=\\\{\}]
{\color{incolor}In [{\color{incolor} }]:} \PY{n}{train\PYZus{}}x} \PY{o}{=}
```

4.1 k-nn classification

4.2 The winning methods

4.2.1 Neural networks

4.2.2 Regularized greedy forest

4.2.3 XGBoost

5 Results on the Kaggle data

5.1 k-nn classification

5.2 Comparision of all methods

6 Discussion

List of Figures

List of Tables