CS281 Status Update

Scalable Signal Region Identification with applications to the ATLAS $W^{\pm}W^{\pm}W^{\pm}$ Analysis

Nicolò Foppiani, Jonah Philion, Baojia(Tony) Tong

https://github.com/tongbaojia/cs281_mlphys

Draft of the final abstract

In experimental particle physics new phenomena is observed and eventually discovered by looking for statistically significant deviations in the count data over the space of the features which are experimentally measured.

Therefore, this projects aims to develop general strategies to identify signal regions in high background parameter space. This can be naively done by cutting the space in baskets and using the basket with the largest significance. However, thanks to the grown of the computational power, it is nowadays possible to exploit Machine Learning. Several techniques have been studied in order to train classifiers between signal and background, such as Decision Trees and Neural Networks. Furthermore, the problem of maximizing the statistical significance instead of classifying signal and background has been investigated.

The whole project has been developed by studying the simulated dataset of $W^{\pm}W^{\pm}W^{\pm}$ events collected by the ATLAS detector. The results are promising, showing important improvements with respect to the cut-based analysis. These methods will be adopted in the official ATLAS analysis.

Problem Statement In particle physics analysis, there is in general some "signal" process that generates data simultaneously as "background" processes generate similar looking data. Given unlabeled data from all processes, our goal is to determine the likelihood that the signal process exists. To generate data $\{x_i\}_N$ lying in $\mathbb{R}^n$, we associate to the signal process a smooth function $s : \mathbb{R}^n \to \mathbb{R}_{>0}$ and to the background process $b : \mathbb{R}^n \to \mathbb{R}_{>0}$ such that for any volume $V \subset \mathbb{R}^n$, the number of background events in this region $n_{bkg}(V)$ and the number of signal events in this region $n_{sig}(V)$ can be modeled as

$$n_{bkg}(V) \sim \text{Pois}\left(\int_V b\right)$$

$$n_{sig}(V) \sim \text{Pois}\left(\int_V s\right)$$

The data is generated from these Poisson processes. We can then determine the likelihood of the existence of the signal process by choosing regions of our parameter space $\{V_j\}$ then comparing the number of events observed in these regions to the values $\{n_{bkg}(V_i)\}$ and $\{n_{sig}(V_i) + n_{bkg}(V_i)\}$. For the case of large N, we can approximate the Poisson distribution as Gaussian. For a region with O observed events, the significance is then

$$Significance \approx \frac{O - n_{bkg}(V)}{\sqrt{n_{bkg}(V)}} \tag{1}$$

We have simplified our problem to determining the set $\{V_i\}$ of regions of parameter space which will maximize the definition of significance in Equation 1.

It is noteworthy to point out that this problem is not trivial. In fact, the simplest approach is to try to classify signal and background: this means to estimate the probability to be signal for each value of the experimental features. Regions with larger fractions of signal events will be more likely to be significant. However this is not always the case, and a proper optimization of the significance should be done.

A second caveat is due to the uncertainty on the significance: we want to identify regions in which the significance is large with small fluctuations. For instance, optimizing the significance might bring to rely on regions in which we have one expected signal event and 10^{-3} background event: this region would be very significant but not reliable, since the *uncertainty* on the significance will be large.

Baselines Since our goal is to determine volumes of parameter space which maximize the significance, one baseline is to take the volume to be the entire parameter space. The metrics for this approach are summarized in table 1.

In particle physics a significance greater than 3σ is required by the community to claim an observation, and a significance larger than 5σ in order to claim a discovery. The observed significance in our dataset is 0.517σ: this is much smaller than what request for observing this rare physical process. Thus, this puts a strong motivation in applying new techniques in order to enlarge this value.

A description of the data used in this analysis is shown in table 2. This table illustrates the various features which have been selected, stressing also if they are continuous or categorical variables. In the end, it will also be of interest to try to reduce the variables to the subset of the most significant ones.

	n_{sig}	n_{bkg}	$n_{sig}/\sqrt{n_{bkg}}$
No Cuts	47.09	8288.18	0.517σ

Table 1: Baseline measurements. n_{sig} is the weighted sum of all signal events, n_{bkg} is the weighted sum of all background events, and $S/\sqrt{B}$ is our metric for the significance. The objective is to maximize the significance, resulting in some value larger than the 3σ limit.

Approaches

1. **Physics Motivated Cuts** Cuts from the most recent analysis paper. This will be the baseline that we try to achieve and beat.

2. **Naive Classification** The first approach is to train a classifier $x \to y$ where x is a list of kinematic and categorical variables and y is the binary label: 0 for background and 1 for signal. The loss function used in this case is the cross entropy, weighted with the relative weights of the events, in order to obtain the correct mixture of events predicted by the theory.
   Three approaches has been pursued:
   - Logistic regression (using Pytorch)
   - Neural network classifier (using Pytorch)
   - Boosted decision trees (using XGBoost)

Variable Name	Description	Type
j^0_m	Invariant mass of the most energetic jet	continuous
$j^0_{p_T}$	Transverse momentum of the most energetic jet	continuous
j^0_η	Pseudorapidity of the most energetic jet	continuous
j^0_ϕ	Azimuthal angle of the most energetic jet	continuous
l^0_m	Invariant mass of the highest p_T lepton	continuous
$l^0_{p_T}$	Transverse momentum of the highest p_T energetic lepton	continuous
l^0_η	Pseudorapidity of the highest p_T lepton	continuous
l^0_ϕ	Azimuthal angle of the highest p_T lepton	continuous
l^0_c	Charge of the highest p_T lepton	categorical
l^0_{isEl}	1 if the most energetic lepton is an electron, 0 for muon	categorical
l^1_m	Invariant mass of the second-to-most energetic lepton	continuous
$l^1_{p_T}$	Transverse momentum of the second-to-most energetic lepton	continuous
l^1_η	Pseudorapidity of the second-to-most energetic lepton	continuous
l^1_ϕ	Azimuthal angle of the second-to-most energetic lepton	continuous
l^1_c	Charge of the second-to-most energetic lepton	categorical
l^1_{isEl}	1 if the second-to-most energetic lepton is an electron, 0 for muon	categorical
l^2_m	Invariant mass of the third-most energetic lepton	continuous
$l^2_{p_T}$	Transverse momentum of the third-most energetic lepton	continuous
l^2_η	Pseudorapidity of the third-most energetic lepton	continuous
l^2_ϕ	Azimuthal angle of the third-most energetic lepton	continuous
l^2_c	Charge of the third-to-most energetic lepton	categorical
l^2_{isEl}	1 if the third-to-most energetic lepton is an electron, 0 for muon	categorical
met_{p_T}	Transverse momentum of the missing transverse energy	continuous
met_{phi}	Azimuthal angle of the missing transverse energy	continuous
$weight$	event weight, normalization to the correct data size, and kinematic corrections	continuous
cl	The physics process that generated event.	categorical
is_{sig}	1 if signal, 0 if background.	categorical

Table 2: Feature Space. We measure cl, is_{sig}, and $weight$ only in simulation. We measure all other variables in real data and in simulation. Our general approach to train classifiers to predict is_{sig} given the jet and lepton features.

3. **Category Specific Classification** We factor the probability that an event is a signal as

$$p(event.is_sig == 1) = p(event.isSig == 1|x_{categorical}) * p(event.isSig == 1|x_{continuous})$$

We model the first probability as a categorical distribution with probabilities given by

$$p(event.isSig == 1|x_{categorical}) = \frac{\sum 1_{category,signal} w}{1_{category} w}$$

We model the second probability using XGBClassifier on a small parameter grid with 3-fold cross validation.
At test time, for every event, we calculate the probabilities $p(event.is_sig == 1)$. We then make a histogram of p and evaluate the significance on a combination of these bins. We calculated the significance for train and test data: they are shown in figure 1 and figure 2, respectively. The raw values are shown in table 3.

	n_{bkg}	n_{sig}	$n_{sig}/\sqrt{n_{bkg}}$
train	44.271	10.995	1.652σ
test	41.478	10.994	1.707σ

Table 3: Category Specific Classification Results. The significance is 328% better than the baseline. We will calculate the variance on this test error with bootstrapping.

4. **Loss function containing significance** Optimizing the crossentropy lets us make inference on the probability to be signal for each value of the features. However, the goal is not to have the greatest possible separation in two classes, but to obtain the highest significance for the discovery. The problem of optimizing an analysis in order to maximize the significance is hard and few studies are present in the scientific literature.
The most naive approach is to make the significance differentiable and optimize it as a loss function. This has not led to any exceptional result so far. Additionally we are also trying to implement a more robust loss function, which we can maximize obtaining the highest significance.

It is expected to summarize together the results by the end of November.

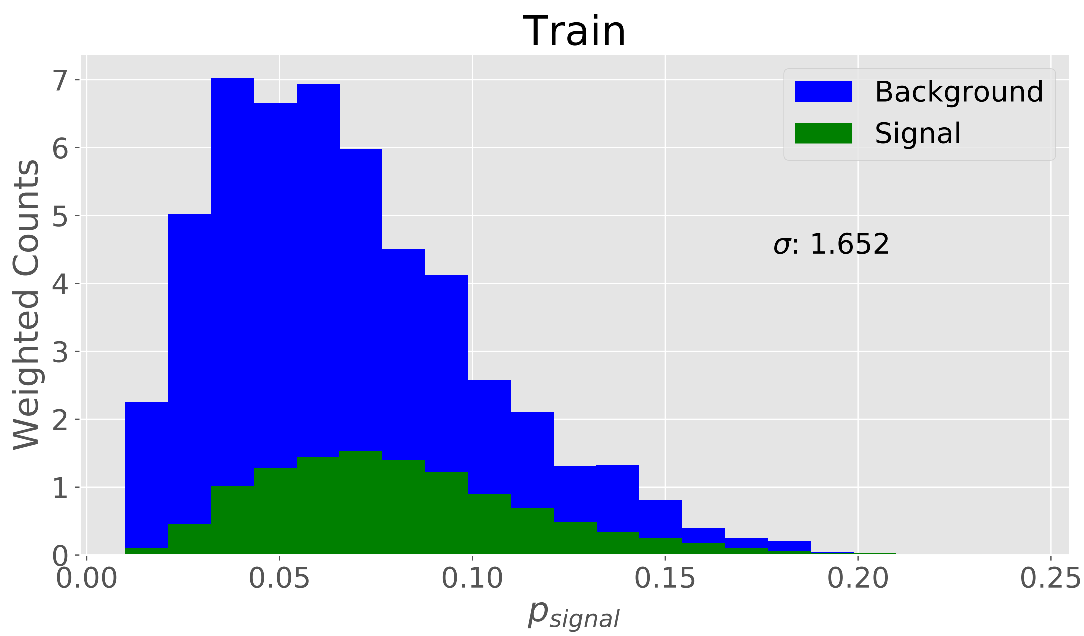

Figure 1: Category Specific Classification. Histograms of estimated probabilities for all events. Only events with $p > .01$ are included in this plot because the background dominates in the region $p < .01$. The significance is computed over bins with $p > 0.0322$. We make this decision based on the training set distribution only.

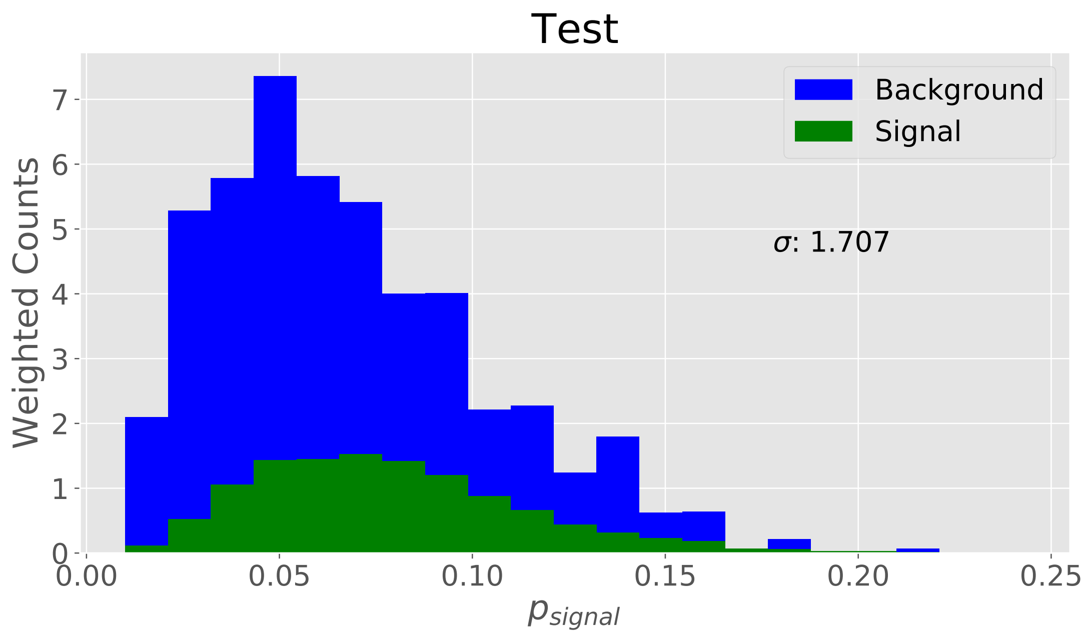

Figure 2: Category Specific Classification. The equivalent graph to 1 but over the test set. The significance is about the same in the test set as it is in the train set. We'll have to run the algorithm on different random train/test partitions (bootstrapping) to get a sense of the variance of this test σ.