TRISEP ML tutorials

Robert Braun, Justus Beisenkötter, Cornelis Mommers

B1

TRISEP ML tutorials

Robert Braun, Justus Beisenkötter, Cornelis Mommers

Beginner project

B1

We completed...

- Intro notebooks
  - Tensor broadcasting and graph back propagation
  - Deep learning building blocks
  - Fundamentals in Numpy

We will discuss these:

- Beginner Notebooks
  - Exercise 1
  - Exercise 2

Exercise 1: Exploring the dataset

The data for the exercise are stored in `/fast_scratch_1/TRISEP_data/BeginnerTutorial/dataWW_d1.root`. This is created from ATLAS Open Data. The dataset is composed of signal and background events for a Higgs analysis decaying to two W bosons ($H \to W^+W^-$). Inside this dataset there are many different variables that we want to explore before starting the classification task. For this task, have a look at `scripts/explore.py` and execute it with

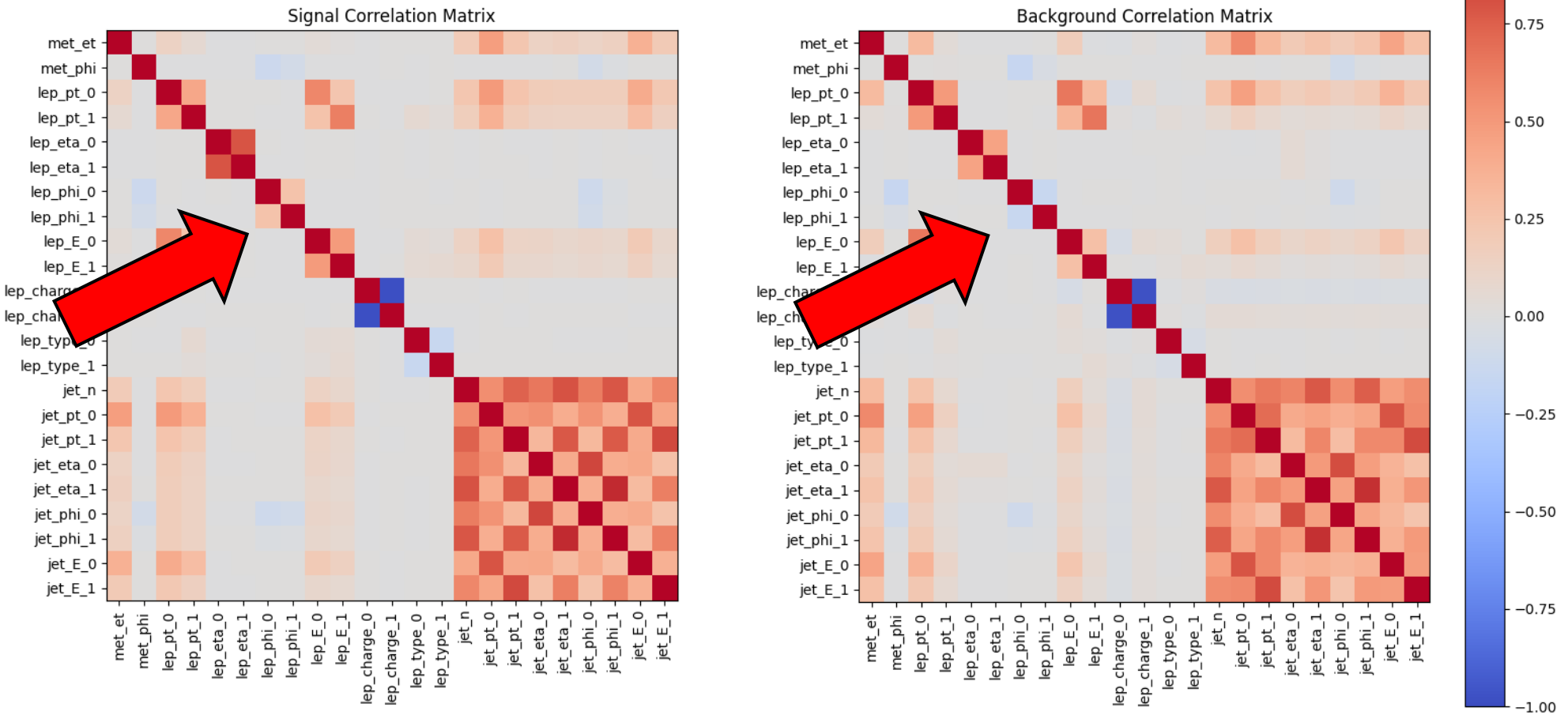
```
python scripts/explore.py
```

This will produce 2 figures inside the folder `figures/`. Have a look at them and try to answer to these:

- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

`lep_phi_0`
`lep_phi_1`
Signal: pos. cor.
BG: neg. cor.



- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

```
np.max(np.abs(corr_signal - corr_bg))
```

```
0.3973703340727768
```

		lep_phi_1
		⋮
lep_phi_0	•••	0.397370

- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

```
np.max(np.abs(corr_signal - corr_bg))
```

```
0.3973703340727768
```

lep_phi_1

⋮

lep_phi_0 ••• 0.397370

- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

```
np.max(np.abs(corr_signal - corr_bg))
```

```
0.3973703340727768
```

		lep_phi_1
		⋮
lep_phi_0	•••	0.397370

But why?

Why is `lep_phi` a good discriminator?

- Helicity conservation & chiral nature of *W* vertex imply that (at *E* >> *m*) positive correlation of phi angles expected
- Here: no time, for details see

Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector

The ATLAS Collaboration

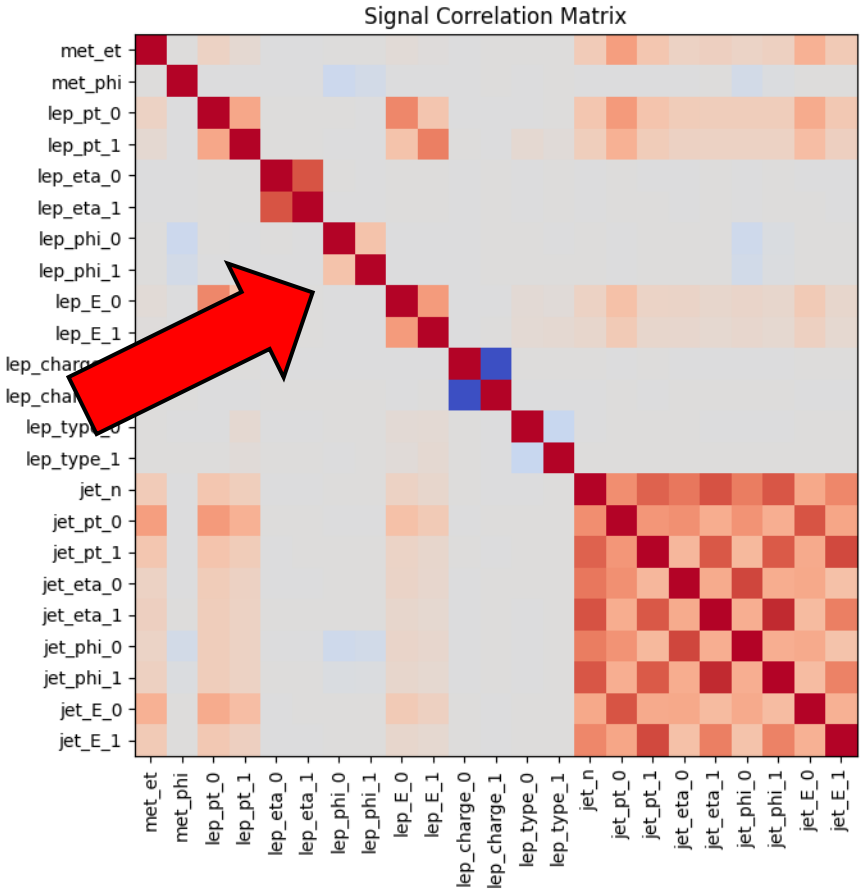
Abstract

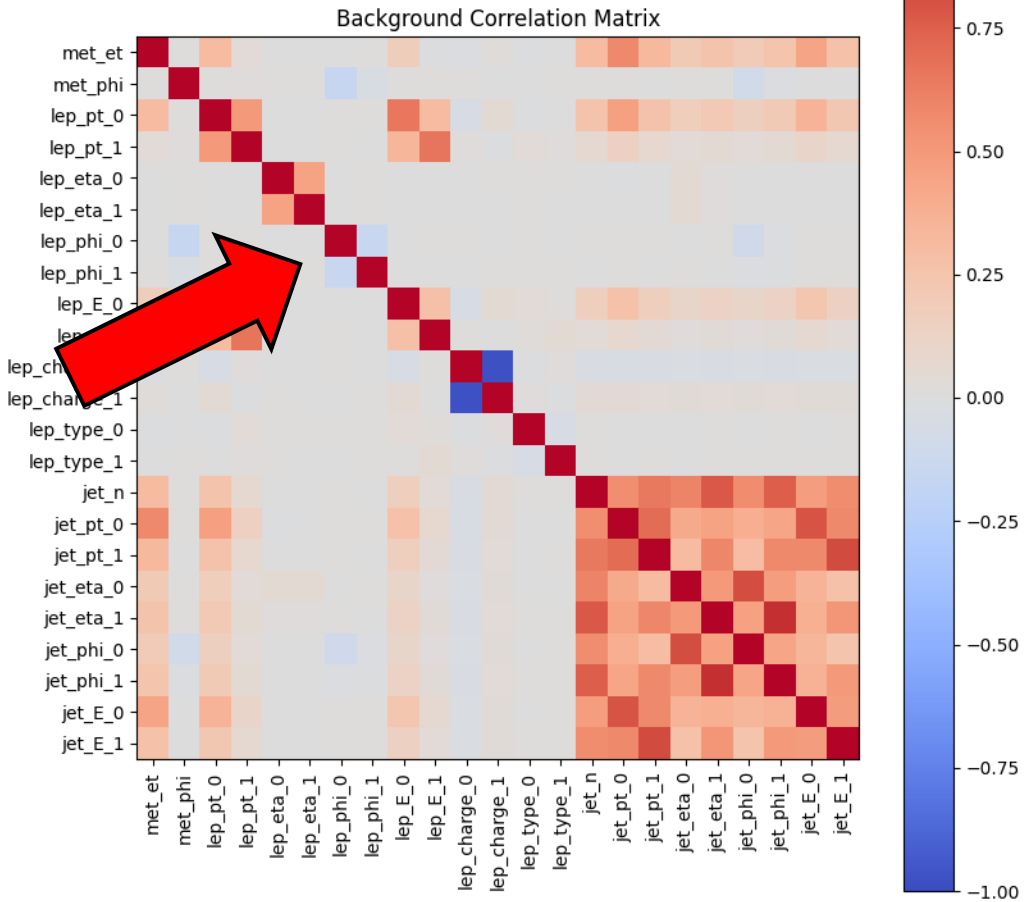
We report the observation of Higgs boson decays to WW^* based on an excess over background of 6.1 standard deviations in the dilepton final state, where the Standard Model expectation is 5.8 standard deviations. Evidence for the vector-boson fusion (VBF) production process is obtained with a significance of 3.2 standard deviations. The results are obtained from a data sample corresponding

- Do you see which variables provide the highest discriminating power between background and signal?
- How are variables correlated?
- Are there variables that are probably not helping with the classification? If so, which ones?
- Try to add more variables to the plots.

`lep_phi_0`
`lep_phi_1`
Signal: pos. cor.
BG: neg. cor.

Variables w/
`|Sig-BG| < ε`
not helpful





Exercise 2: Multi-layer perceptron (MLP) for Higgs classification

Classification is one of the most common tasks in High Energy Physics. Its goal is to combine various inputs from data to obtain a numbers later used to classify data in different categories. A very standard task is the background vs signal classification, where each event is assigned with a class number 0 (background) and 1 (signal).

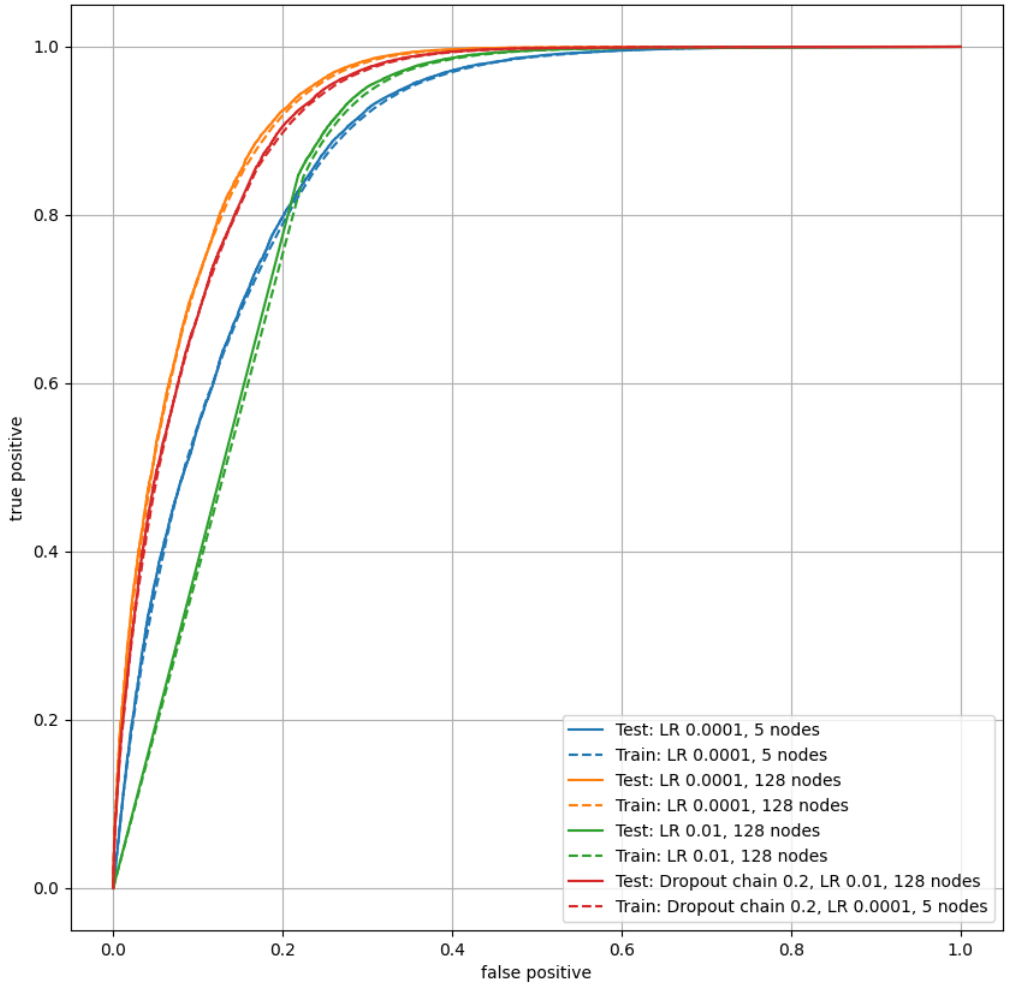
In this tutorial we will try to classify $H \to W^+W^-$ signals from background in 2-lepton final states.

A few exercises:

1. Try to increase the number of nodes in the hidden layers to 128 inside trisepmltutorial/mlp.py. Is the network doing better for classification after training? Or worse?
2. Exercise: try to increase the learning rate. Is the learning time improving? Is the precision better?
3. Exercise: try to prevent overtraining by adding dropout layers (see nn.Dropout).

1. Try to increase the number of nodes in the hidden layers to 128 inside trisepmltutorial/mlp.py. Is the network doing better for classification after training? Or worse?
2. Exercise: try to increase the learning rate. Is the learning time improving? Is the precision better?

1. **More hidden layers: better performance**
2. **Higher learning rate: higher true positive**

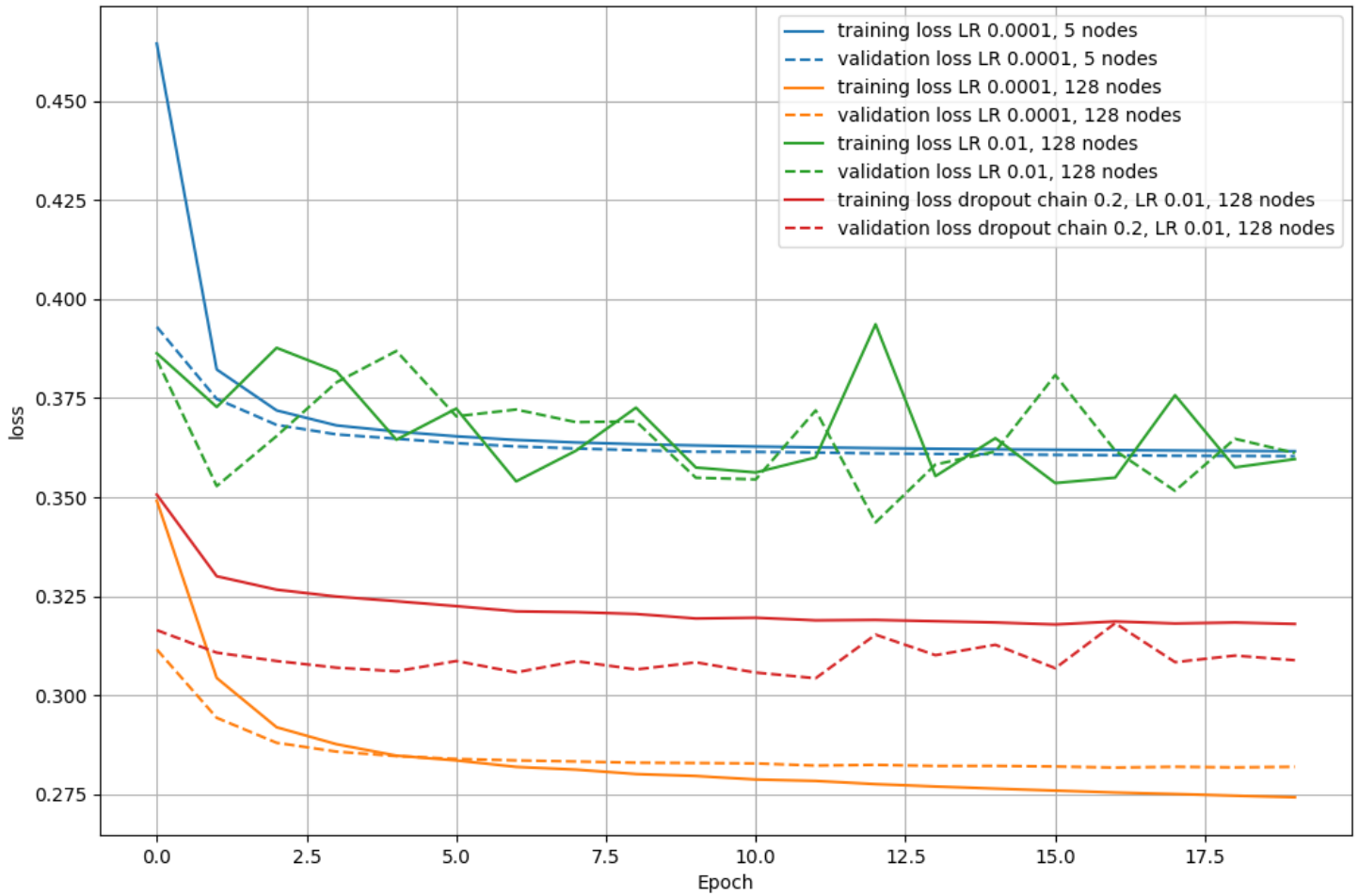


3. Exercise: try to prevent overtraining by adding dropout layers (see nn.Dropout).

3. Overtraining seen by comparing training loss with validation loss.

Decreasing training loss while validation loss does not decrease is evidence of too many epochs, since model does not improve on 'validation check'.

Dropout cures this problem.



Thank you for your attention!