Weakly supervised classifiers
learning from data and proportions

L. Dery (Stanford), B. Nachman (LBNL), F. Rubbo (SLAC), A. Schwartzman (SLAC)
LHC detectors as cameras

The LHC experiments are $O(100)$ Megapixel 3D fast cameras —> High resolution “pictures” of proton-proton collisions.

E.g. Jets are recorded as densely packed tracks and calorimeter “images”.

Broad effort aiming at outperforming Physics-motivated feature extraction by using low-level inputs (e.g. calorimeter “pixels”) for ML algorithms.
Jet classification example

RNN

CNN

Occupancy

Fisher-Jet Output
Jet classification example
Learning from simulation vs learning from data

• Modeling of multi-dimensional soft QCD features (e.g. $n_{\text{track}}, w_{\text{track}}$) is challenging for MC.

• Expect further strain at higher dimensionality (e.g. images with thousands of pixels!)
Learn directly from unlabeled data!

Weakly supervised classifier trained without using labels

Traditional fully supervised classifier
Traditional full supervision

Labeled training set (“simulation”)

Classification

\[ f_{\text{full}} = \arg \min_{f: \mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell(f'(x_i) - t_i) \]

\[ f_{\text{full}}(\text{apple}) = 0.97 \]
Weak supervision

unlabeled training data

average composition for each barrel

\[ f_{\text{weak}} = \arg\min_{f': \mathbb{R}^n \to [0,1]} \ell \left( \frac{1}{N} \sum_{i=1}^{N} \frac{f'(x_i)}{y} \right) \]

Classification \( f_{\text{weak}}(\text{apple}) = 0.97 \)
Weak supervision - analytically

**unlabeled data sample A**

\[ y_A = 0.1 \]

**unlabeled data sample B**

\[ y_B = 0.3 \]

\[ h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i} \]

\[ h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i} \]

- Given two independent unlabeled data samples, and the corresponding proportion of signal, we can extract the signal and background distributions.
Weak supervision - analytically

Given two independent unlabeled data samples, and the corresponding proportion of signal, we can extract the signal and background distributions.

\[ h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i} \]
\[ h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i} \]

\( y_A = 0.1 \)
\( y_B = 0.3 \)
Weak supervision

- The analytic approach requires binning and becomes quickly unmanageable as the feature space grows.

- ML approach directly looks for discriminant, without extracting explicitly n-dimensional feature distributions for S and B.

\[
f_{\text{full}} = \arg\min_{f': \mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell(f'(x_i) - t_i)\\
\]

\[
f_{\text{weak}} = \arg\min_{f': \mathbb{R}^n \to [0,1]} \ell\left(\frac{\sum_{i=1}^{N} f'(x_i)}{N} - y\right)
\]

Same discrimination as fully supervised
Weak supervision - q/g tagging

\[ |\eta_{j1}| - |\eta_{j2}| \text{ in dijet events} \]

Leverage precise description of ME and PDF (MC/theory) to extract discrimination from soft QCD features (from data!)

Each bin is a “barrel” of jets with known proportion

1/4 quarks
Conclusion

- **Weak supervision** is a new paradigm the **class proportions** in high-level observables in order to use **unlabeled data** to extract **discriminating information** from poorly modeled or unknown **low-level observables**.


SLAC-PUB-13402
References

• Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s}=7$ TeV with the ATLAS detector - https://arxiv.org/abs/1405.6583
• Weak supervision allows training directly on data
• Learns only real features, from being exposed to discriminant features in data.

**Same performance as ideal classifier, trained on labeled data**

**1405.6583**
Stability