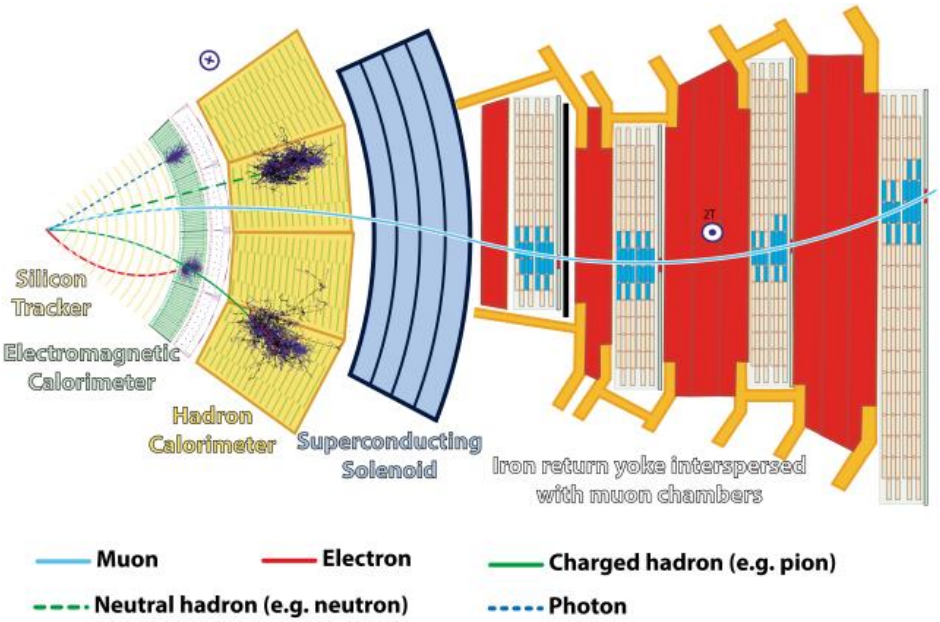
Multivariate analysis for the $W \longrightarrow \pi\gamma$ search

Jeske Dioquino
University of California, Davis
Mentor: Mario Pelliccioni

Outline

1. CMS - the experiment and my part in it
2. $W \longrightarrow \pi\gamma$ events
3. ROOT TMVA package
4. Training with Monte Carlo (MC) signal/background events
5. BDT score and event selection
6. Conclusions and next steps

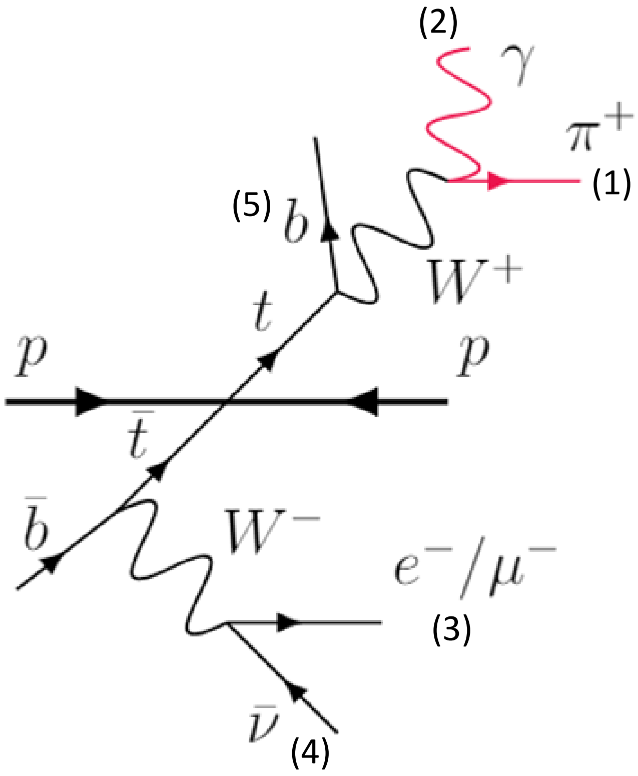
CMS - the experiment and my part in it



- ❖ General-purpose detector with different detection layers
- ❖ Large international collaboration
- ❖ EP-UCM: Experimental Physics department, CMS users group

$W \longrightarrow \pi\gamma$ events

- Very rare – BR < 10^{-5}
- Multiple variables to consider
  - p_T^{π} (1)
  - E_T^{γ} (2)
  - $p_T^{e/\mu}$ (3)
  - event missing energy (4)
  - nBjets (p_T>25GeV) (5)
  - π relative isolation
  - e/μ relative isolation
  - $\Delta\phi(\ell,\pi)$



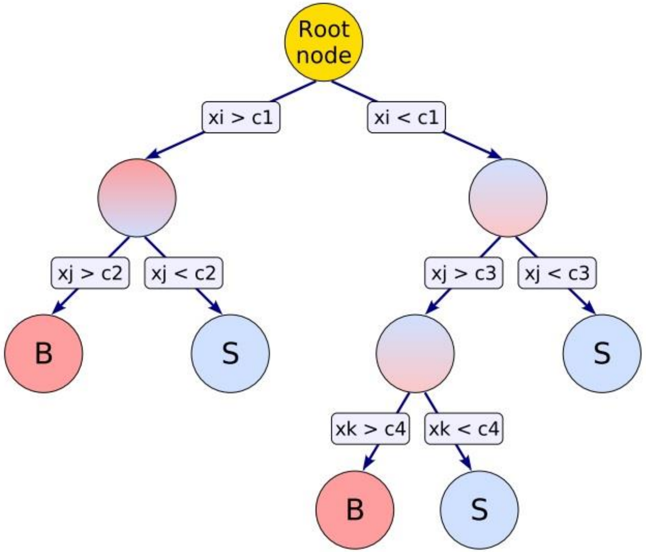
ROOT TMVA package

from TMVA Home

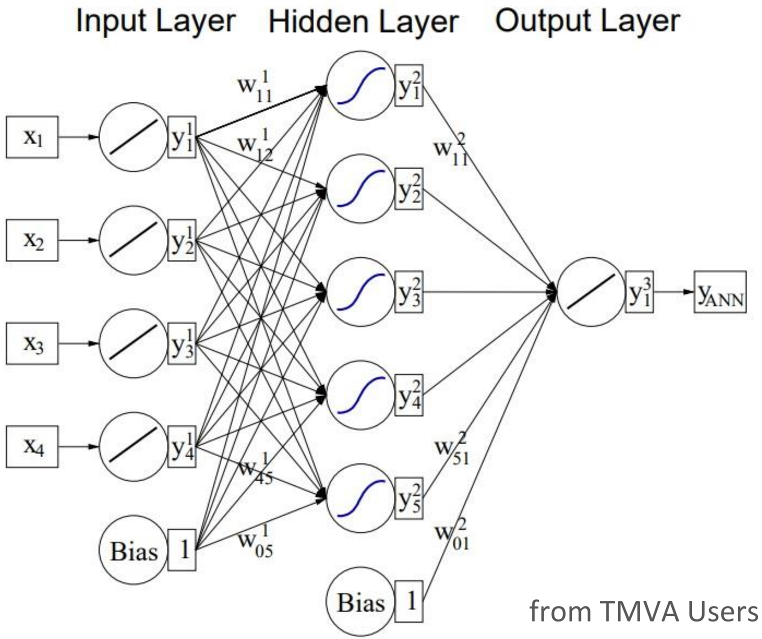
- TMVA = Toolkit for MultiVariate Analysis
- Includes supervised learning algorithms for classification and regression
- My task was to determine which classification method best discriminates between signal and background for $W \longrightarrow \pi\gamma$, and which of the variables matter
- For training and testing the MVA methods, I was given Monte Carlo signal and background events in TTrees containing the different input variables

Training with MC signal/background events

Boosted decision trees (BDT)



MLP neural network (NN)



from TMVA Users Guide

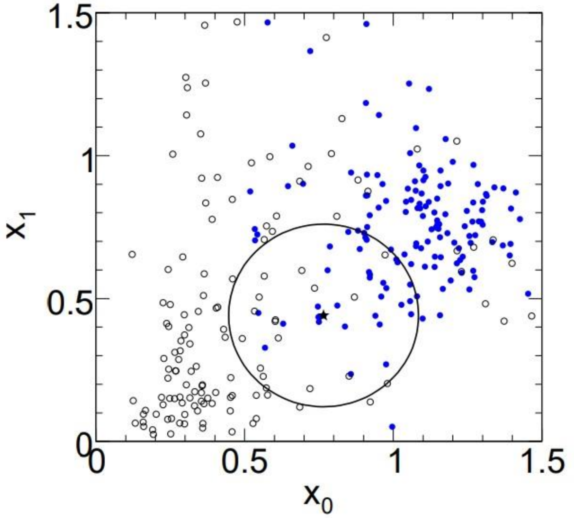
Training with MC signal/background events

Projective likelihood estimator (PDE)

$$y_{\mathcal{L}}(i) = \frac{\mathcal{L}_S(i)}{\mathcal{L}_S(i) + \mathcal{L}_B(i)}$$

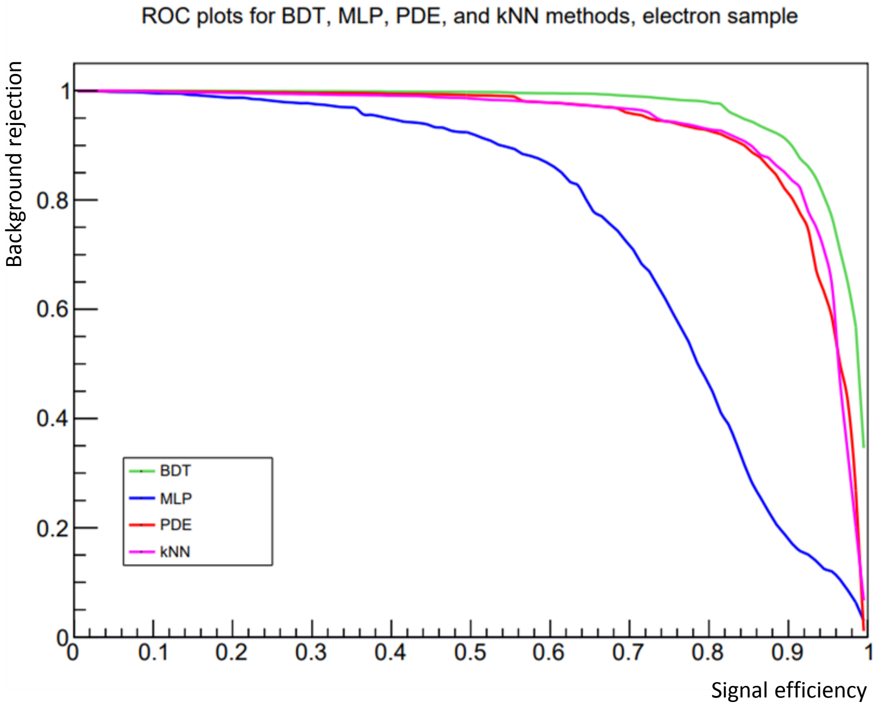
$$\mathcal{L}_{S(B)}(i) = \prod_{k=1}^{n_{\text{var}}} p_{S(B),k}(x_k(i))$$

k-Nearest Neighbour (k-NN)



from TMVA Users Guide

Training with MC signal/background events



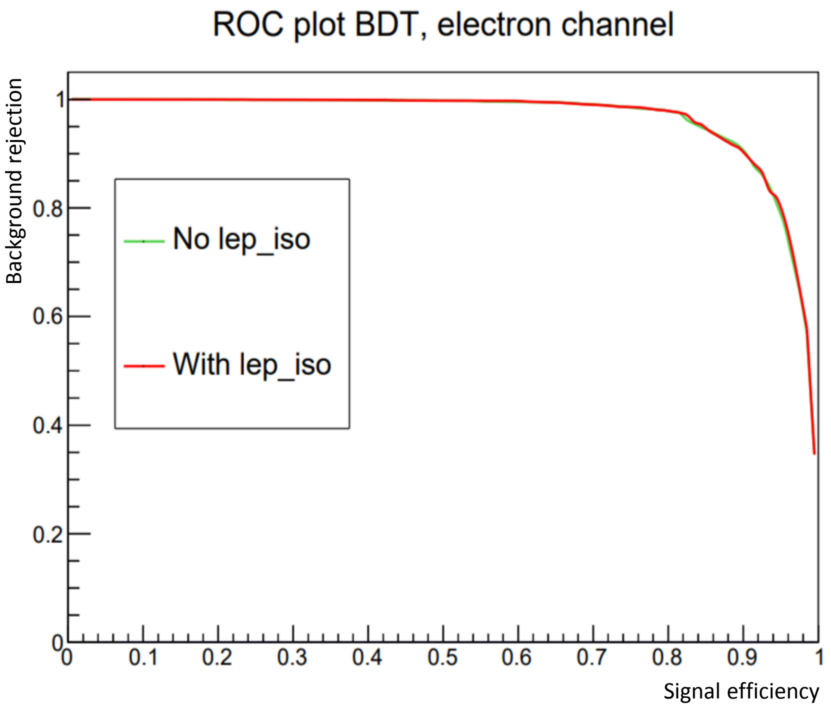
BDT clearly performs the best for this selection

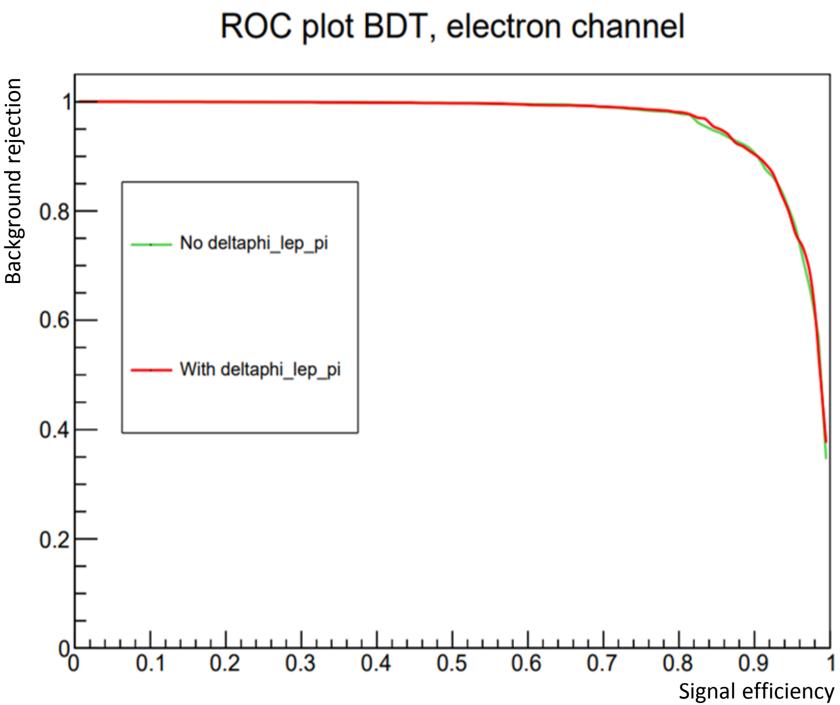
MLP performance is limited by our computing power

PDE and kNN perform similarly, as expected, as they both use probability density functions (PDFs)

- ❖ Didn't use all the input variables

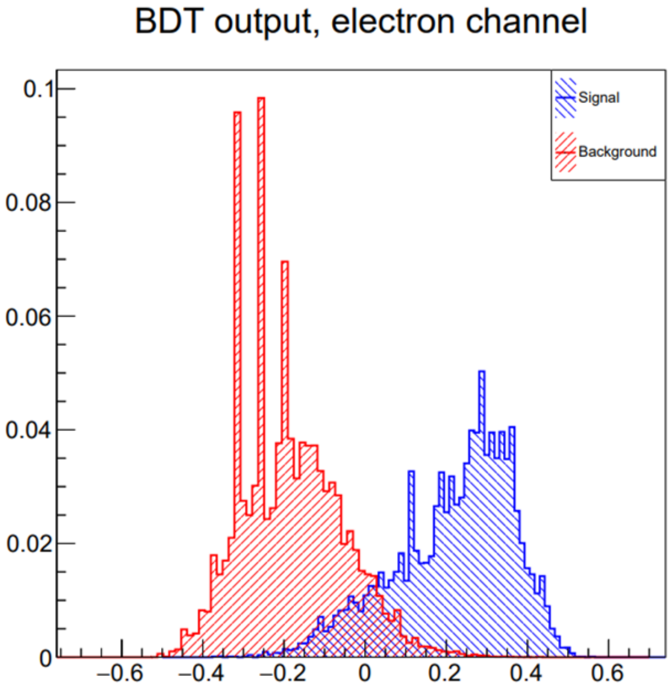
Training with MC signal/background events





BDT score and event selection

- BDT gives each event a score between -1 and 1 based on how “signal-like” it is
- Based on our MC sample’s BDT score distribution we can choose an appropriate cut that gives us more signal and fewer background events
- Fluctuations in background distribution due to some low background statistics



*normalized

Conclusions and next steps

- For the BDT the variables $\Delta\phi(\ell,\pi)$ and e/μ relative isolation did not have much effect on performance → can make the training more robust by excluding them
- For the given sample of MC events the BDT method is most effective at discriminating signal and background
  - PDE and kNN do fine but have intrinsic limits
  - MLP is limited by computing resources
- After this work, next question is, what is the cut on BDT score that maximizes significance?

$$Sig = \frac{N_S \cdot \varepsilon_S}{\sqrt{N_S \cdot \varepsilon_S + N_B \cdot \varepsilon_B}}$$


Le Cry
Les Arolles