НИЯУ МИФИ

Применение методов машинного обучения для идентификации струй, образованных W-бозоном

Научный руководитель:

Мягков Алексей Григорьевич

Студент:

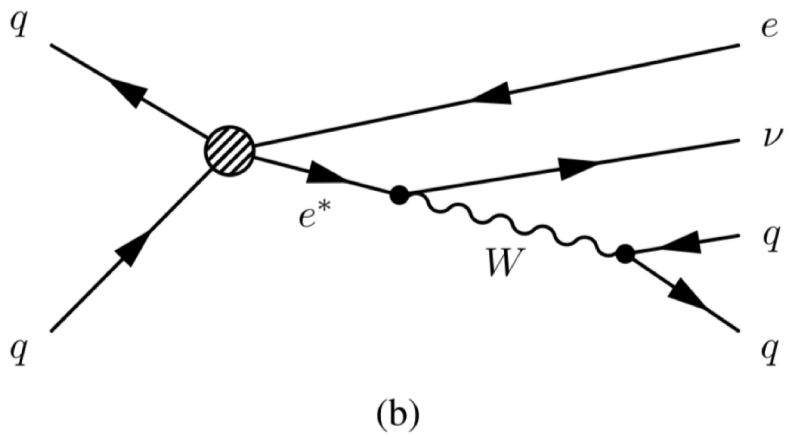
Ван Алина Маошэновна

28.12.2023

Мотивация

Проблемы Стандартной модели:

- Скрытая масса
- Проблема иерархии масс и структуры поколений
- Темная энергия и т.д.



(b)

Пример: Поиск возбужденного лептона с последующим распадом через калибровочный бозон

Идентификация толстых струй, образованных W-бозонами, распавшимися по адронной моде, является одним из главных составляющих этапов анализа данных с экспериментов по поиску новой физики.

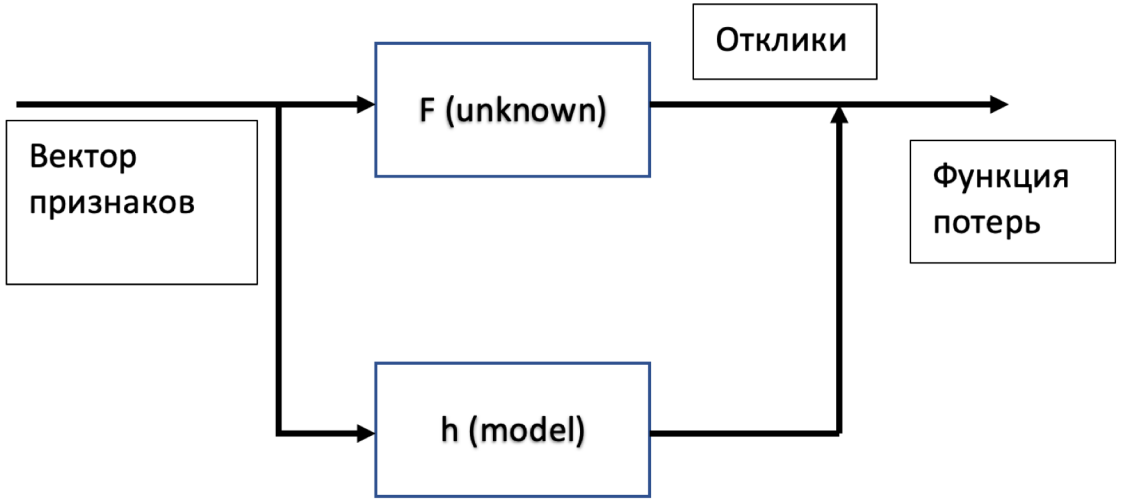
Цель и задачи

Цель: использование методов машинного обучения для решения задачи идентификации толстых струй, образованных W-бозоном

В соответствии с поставленной целью задачами данной работы были:

- Ознакомление с основными алгоритмами машинного обучения для решения задачи бинарной классификации;
- Выбор дискриминирующих переменных (признаков) для объектов;
- Формирование сигнального и фонового деревьев для данной задачи;
- Обучение и тестирование моделей;
- Резюме результатов.

Машинное обучение с учителем



- Машинное обучение – это область прикладной математики, изучающая методы решения задач с использованием обучающих данных.

На вход алгоритма подаются размеченные данные вида (матрица признаков; отклик)

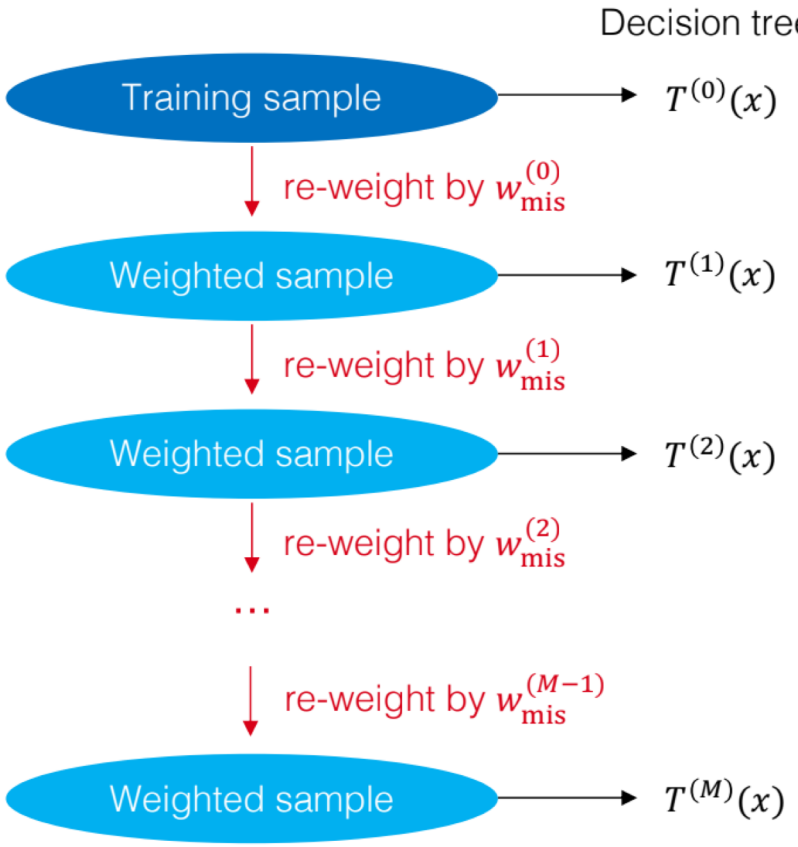
Эти данные называются **обучающей выборкой**, они используются для настройки модели.

Тестовая выборка используется для финальной оценки качества модели.

- Цель машинного обучения с учителем состоит в **предсказании откликов на новых данных**

Методы МО для решения задачи бинарной классификации

AdaBoost



- Задача идентификации частицы – это задача **бинарной классификации**.
- Отклики в данной модели могут принимать только два значения – **сигнал или фон**.

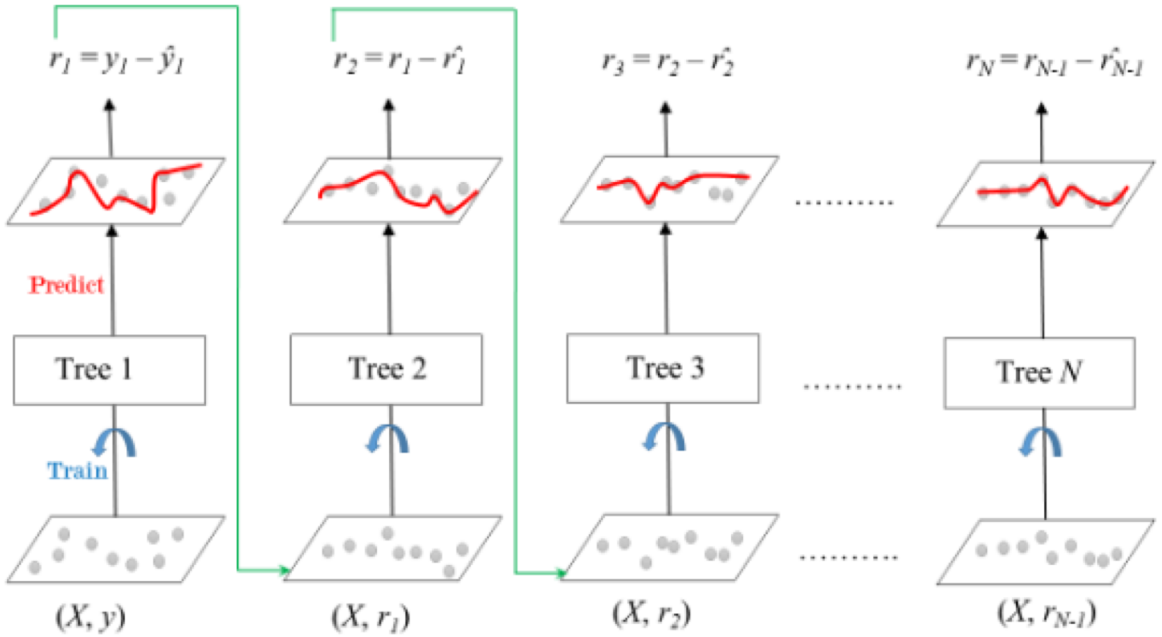
1. Каждое **неглубокое дерево** обучается на случайной подвыборке из обучающих данных;
2. Наблюдениям с **большей ошибкой** приписывают **больший вес**. Вес наблюдения определяет вероятность попадания этого события в обучающую выборку для последующего дерева;
3. Слабым моделям в ансамбле приписывается вес в соответствии со степенью доверия к ним;
4. Выход ансамбля деревьев вычисляется как **усреднение выходов** слабых моделей.

Методы МО для решения задачи бинарной классификации

Градиентный бустинг

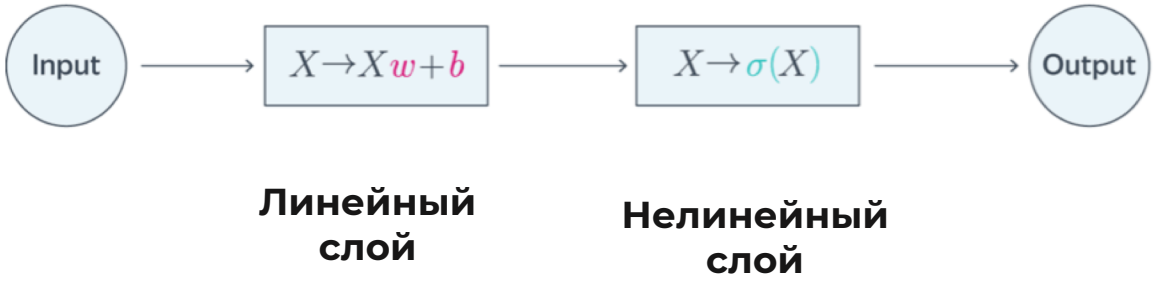
В отличие от адаптивного бустинга, в градиентном бустинге веса для обучающей выборки не обновляются, а вместо этого каждая слабая модель обучается с использованием остаточных ошибок предшествующей модели (псевдоостатках).

Псевдоостатки - это отрицательный градиент функции потерь по выходу модели.



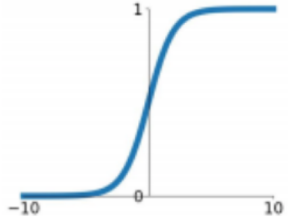
Методы МО для решения задачи бинарной классификации

Нейронная сеть – это сложная дифференцируемая функция, задающая отображение из признакового пространства в пространство ответов, все параметры которой могут настраиваться одновременно и взаимосвязано. Сложную функцию обычно представляют в виде композиции простых функций, которые называют слоями.



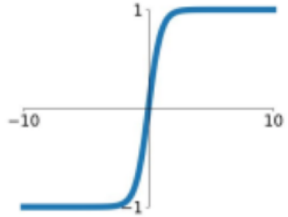
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



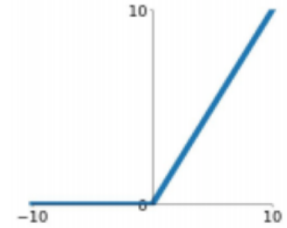
tanh

$$\tanh(x)$$



ReLU

$$\max(0, x)$$



Построение сигнального и фонового деревьев

Сигнальное дерево строится на данных процесса образования пары топ-антитоп

Отбор событий:

- Как минимум одна b-меченная струя
- $p_t > 30$ ГэВ
- $|\eta|_{jet} < 2.5$

Фоновое дерево строится на данных процесса распада Z бозона на электрон-позитронную пару , поставлены ограничения на отсутствие b-меченных струй в событиях .

И на фоновое, и на сигнальные деревья наложены ограничения на поперечный импульс толстой струи и на ее инвариантную массу.

$$60\text{GeV} < m < 110\text{GeV}, p_t > 200\text{GeV}$$

Обучение и тестирование моделей

Используемые методы:

- BDT
- BDTG
- MLP
- DNN

Таблица 1: Гиперпараметры BDT

Гиперпараметр	1	2	3	4	5	6	7	8	9
NTrees	850	850	900	850	850	850	700	700	800
MinNodeSize	2.5%	2.5%	2.5%	1%	2.5%	3%	3%	2.5%	5%
MaxDepth	3	5	3	3	3	3	3	3	3
nCuts	20	20	20	20	40	20	20	20	20

Таблица 2: Гиперпараметры BDTG

Гиперпараметр	1	2	3	4	5	6	7	8	9
NTrees	850	850	900	850	850	850	700	600	800
MinNodeSize	2.5%	2.5%	2.5%	1%	2.5%	3%	3%	3%	5%
MaxDepth	5	5	3	3	3	3	3	3	3
nCuts	500	20	20	20	40	20	20	20	20

Таблица 4: Гиперпараметры DNN

Гиперпараметр	1	2	3	4	5	6	7	8	9
Layout	TANH	RELU	TANH	TANH	TANH	TANH	TANH	TANH	RELU
LearnRate	1e-2	1e-2	1e-2	1e-2	2e-2	2e-2	2e-2	2e-2	1e-2
BatchSize	100	100	100	100	200	200	400	800	800
Regularisator			L2	L2	L2	L2	L2	L2	L2
Drop.Config.	0.5*3	0.5*3	0.5	0.5*3	0.5*3	0	0	0	0

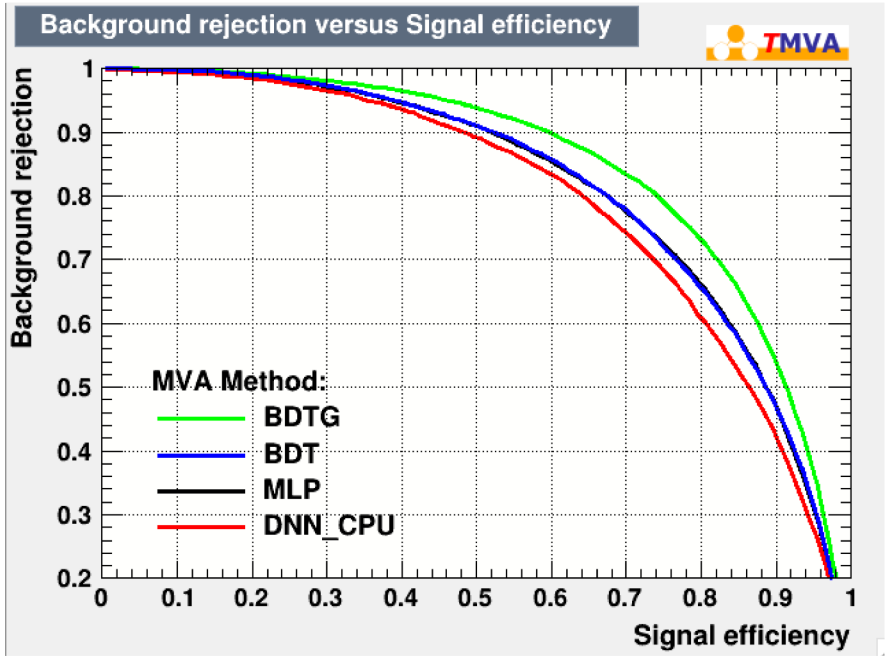
Таблица 3: Гиперпараметры MLP

Гиперпараметр	1	2	3	4	5	6	7	8	9
NeuronType	tanh	ReLU	tanh	tanh	tanh	tanh	tanh	tanh	sigmoid
HiddenLayers	N+6	N+5,N	N+5,N+4	N+5,N+4	N*2,N	N*2,N*2	N*3,N	N*2,N-1	N*2,N-1
NCycles	600	600	600	600	700	700	600	600	600
TestRate	5	5	5	5	5	5	5	5	10

Таблица 5: Значения AUC

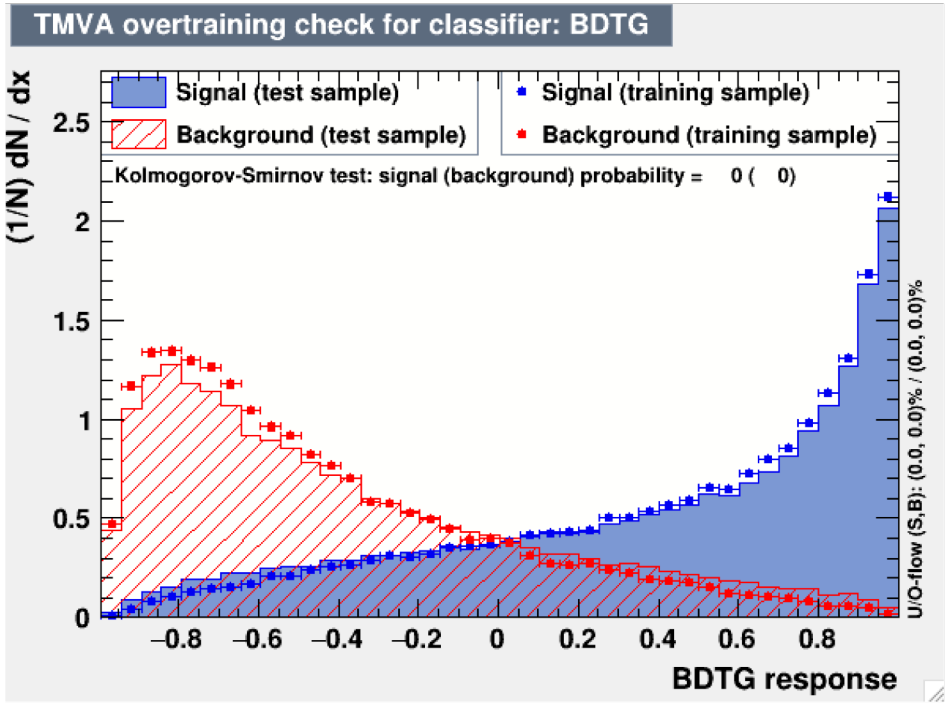
Метод	1	2	3	4	5	6	7	8	9
BDT	0.815	0.824	0.816	0.815	0.815	0.816	0.815	0.814	0.814
BDTG	0.847	0.834	0.822	0.824	0.823	0.821	0.820	0.819	0.811
MLP	0.815	0.815	0.816	0.816	0.816	0.816	0.816	0.816	0.816
DNN	0.794	0.783	0.790	0.777	0.791	0.805	0.808	0.810	0.812

Первый случай



ROC-кривые

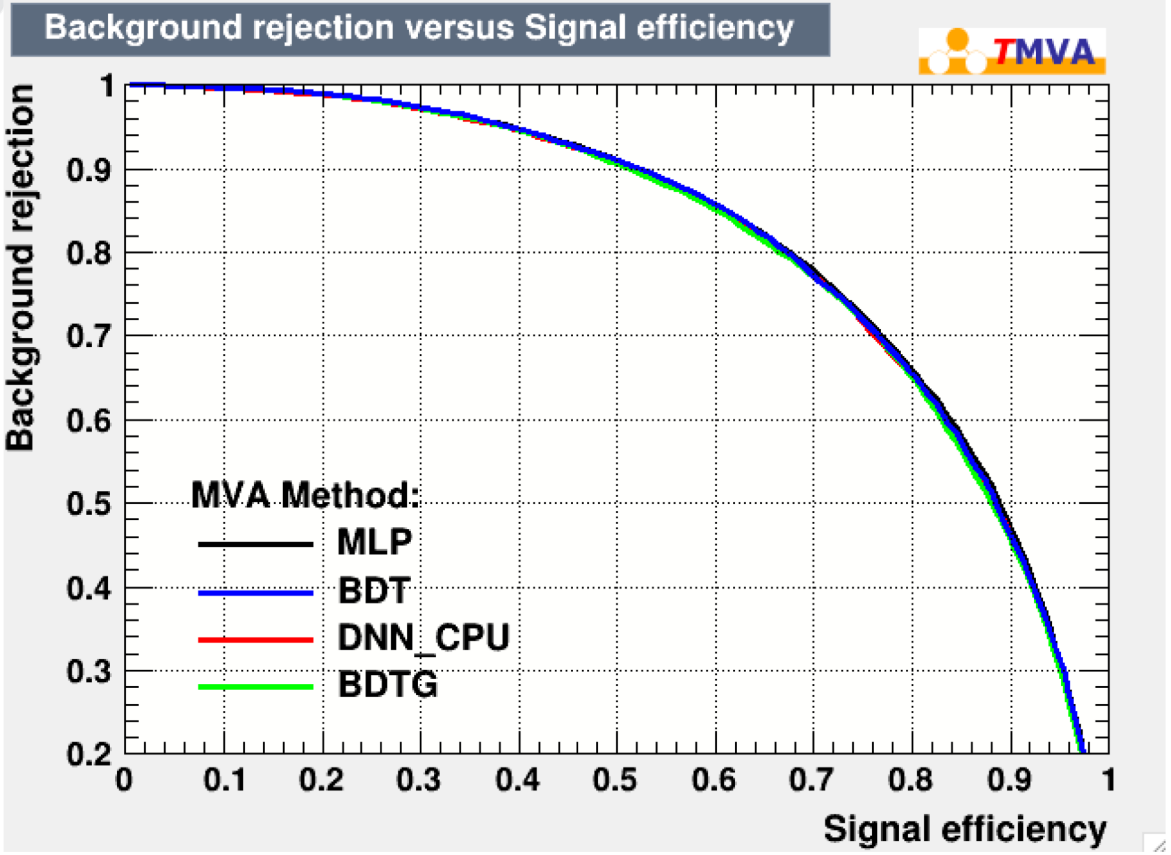
По характеру полученных кривых можно сделать вывод, что классификатор **BDT с градиентным бустингом** лучше разделяет два класса, так как кривая этого метода лежит выше остальных.



Однако при анализе гистограмм с распределением откликов для обучающей и тестовой выборок можно сказать, что при данной настройке гиперпараметров **модель переобучена**.

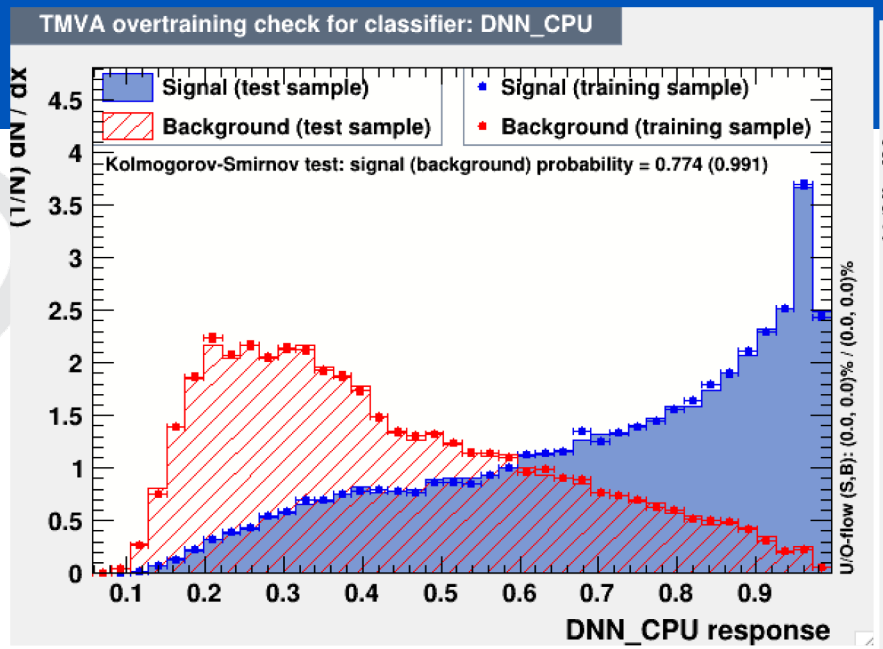
Девятый случай

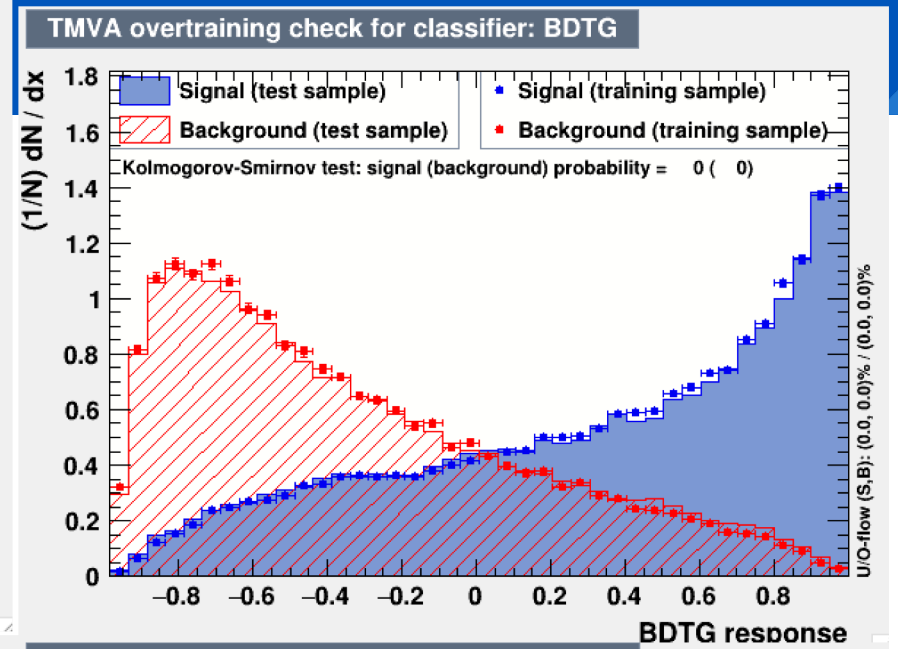
ROC - кривые

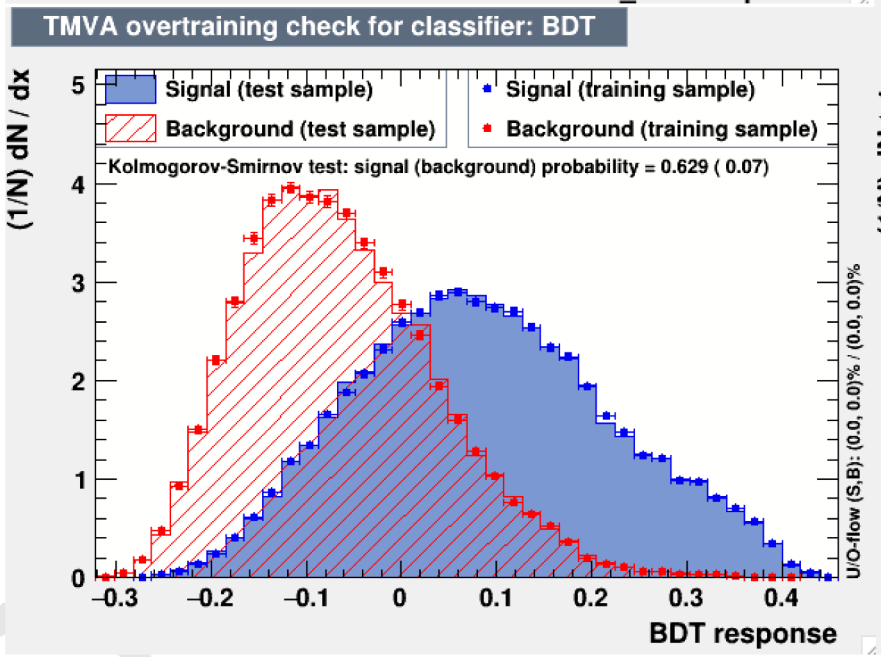


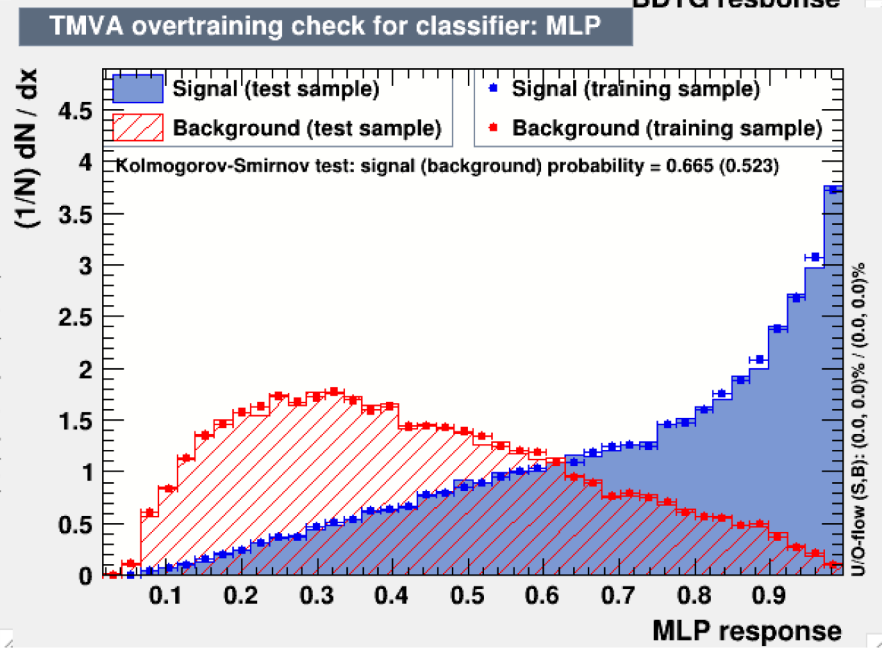
По характеру полученных кривых можно сделать вывод, что результаты классификаторов одинаковые, так как кривые методов совпадают.

Девятый случай









Проверка на переобучение

Для методов тренда на переобучение по полученным распределением не было выявлено.

Заключение

В рамках НИР за семестр проведено ознакомление с основными алгоритмами машинного обучения для решения задачи бинарной классификации; сформированы сигнальные и фоновые деревья , содержащие выбранные дискриминирующие переменные для данной задачи. Проведено обучение и тестирование моделей для разных наборов гиперпараметров.

На основании полученных результатов сделан вывод о том, что BDT с адаптивным бустингом и градиентным бустингом обладают интерпретируемостью и быстрым темпом обучения. В дальнейшей работе будут подробнее рассмотрены алгоритмы обучения глубоких нейронных сетей из-за их лучшей гибкости.

В дальнейшем планируется расширить ассортимент алгоритмов глубокого обучения с использованием других библиотек в Python.