Министерство науки и высшего образования Российской Федерации
Федеральное государственное автомное образовательное учреждение высшего образрвания
«Национальный исследовательский ядерный университет «МИФИ»

УДК 539.12.01

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

ИЗУЧЕНИЕ ПРОГРАММНОГО ПАКЕТА ROOT И ПРИМЕНЕНИЕ ЕГО МЕТОДОВ В РАЗДЕЛЕНИИ ФОНОВЫХ И СИГНАЛЬНЫХ СОБЫТИЙ

Научный руководитель
к.ф-м.н. _____ А. Г. Мягков
Студент _____ А. М. Ван

Москва 2021

Содержание

Перечень сокращений, обозначений и определений

TMVA Toolkit for Multivariate Analysis

BDT Boosted/bagged decision trees

SVM Support vector machine

RuleFit Predictive learning via rule ensembles

ROC Receiver Operation Characteristic

ООП Объектно-Ориентированное Программирование

Введение

Одной из главных задач LHC является поиск новых явлений за рамками Стандартной модели. Несмотря на все свои преимущества, Стандартная модель не дает описания всех известных экспериментальных фактов (темная материя, проблема иерархии и т.д.) . Чтобы обнаружить проявления новой физики, необходимо использовать методы, позволяющие отделить эти проявления от фоновых событий. Поэтому **целью** моей научно-исследовательской работы является изучение программного пакета ROOT и методов многомерного анализа.

1 Основы изучения программного пакета ROOT

Программный пакет ROOT [1] - это объектно-ориентированная среда для обработки и анализа данных, созданная в Европейском центре ядерных исследований. В настоящее время он используется во всех крупных лабораториях ядерной физики и высоких энергий по всему миру для мониторинга, хранения и анализа данных. До ROOT CERN поддерживал свою программную библиотеку PAW, написанную на языке Fortran. В 1994 году была инициирована разработка ROOT сотрудниками CERNa Рене Брюном и Фонсом Рэйдмэйкерсом. В 1995 году они выпустили первую версию программного пакета, реализованного на принципах ООП.

ROOT предоставляет кроссплатформенный интерфейс к графической подсистеме и операционной системе используя механизмы абстракции данных. Частями пользовательского интерфейса являются:

- GUI(Графический интерфейс пользователя);
- CINT(Командный интерпретатор);
- библиотеки, программы (C++ компилятор и интерпретатор);

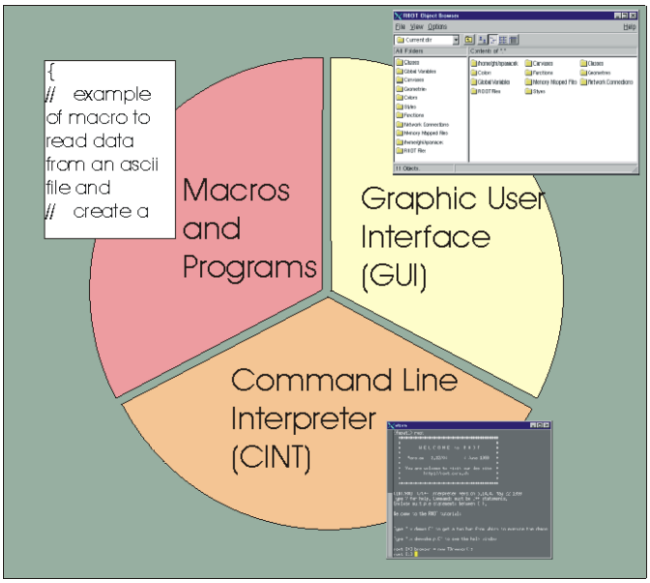

Рис. 1: Three User Interface

Пакеты, включенные в ROOT, содержат:

- стандартные математические функции,
- средства четырехвекторных вычислений,
- инструменты статистического анализа,
- интсрументы линейной алгебры,
- инструменты для создания графиков, гистограмм и т.д.

Ключевой возможностью пакета ROOT является специальный контейнер данных, называемый деревом (Tree), вместе с его подмножествами ветвями (Branch) и листьями (Leaf). Дерево может быть представлено как удобное средство чтения и записи данных в файле. Следующий элемент данных, записанный в файле, может быть получен инкрементированием индекса дерева. Такой подход позволяет избежать проблем с выделением памяти при создании объектов, и даёт возможность дереву выступать в качестве «лёгкого» контейнера при буферизации данных.
ROOT разрабатывался как высокопроизводительная вычислительная библиотека, необходимая для обработки данных Большого Адронного Коллайдера, поток которых достигает нескольких петабайт в год.

1.1 ООП, классы и методы, указатели

Идеологически ООП - это подход к программированию, как к моделированию окружающего мира как совокупности объектов, взаимодействующих друг с другом. Класс — универсальный, комплексный тип данных, состоящий из тематически единого набора «полей» и «методов». Методы - процедуры и функции, связанные с классом. Они определяют действия, которые можно выполнять над объектом такого типа, и которые сам объект может выполнять. Поля - переменных более элементарных типов. Классы образуют иерархию наследования. Наследование — свойство системы, позволяющее описать новый класс на основе уже существующего с частично или полностью заимствованной функциональностью. Класс, от которого производится наследование, называется базовым, родительским или суперклассом. Новый класс — потомком, наследником, дочерним или производным классом. Объект — это конкретный экземпляр, представитель данного класса. Часто в работе с объектами используются указатели.
Пример поинтера:
root[1] TCanvas *MyC = new TCanvas("MyC "Test canvas 1) - здесь создается объект класса TCanvas
root[2] MyC->Divide(2,2) - обращение к методу Divide

1.2 Tutorials

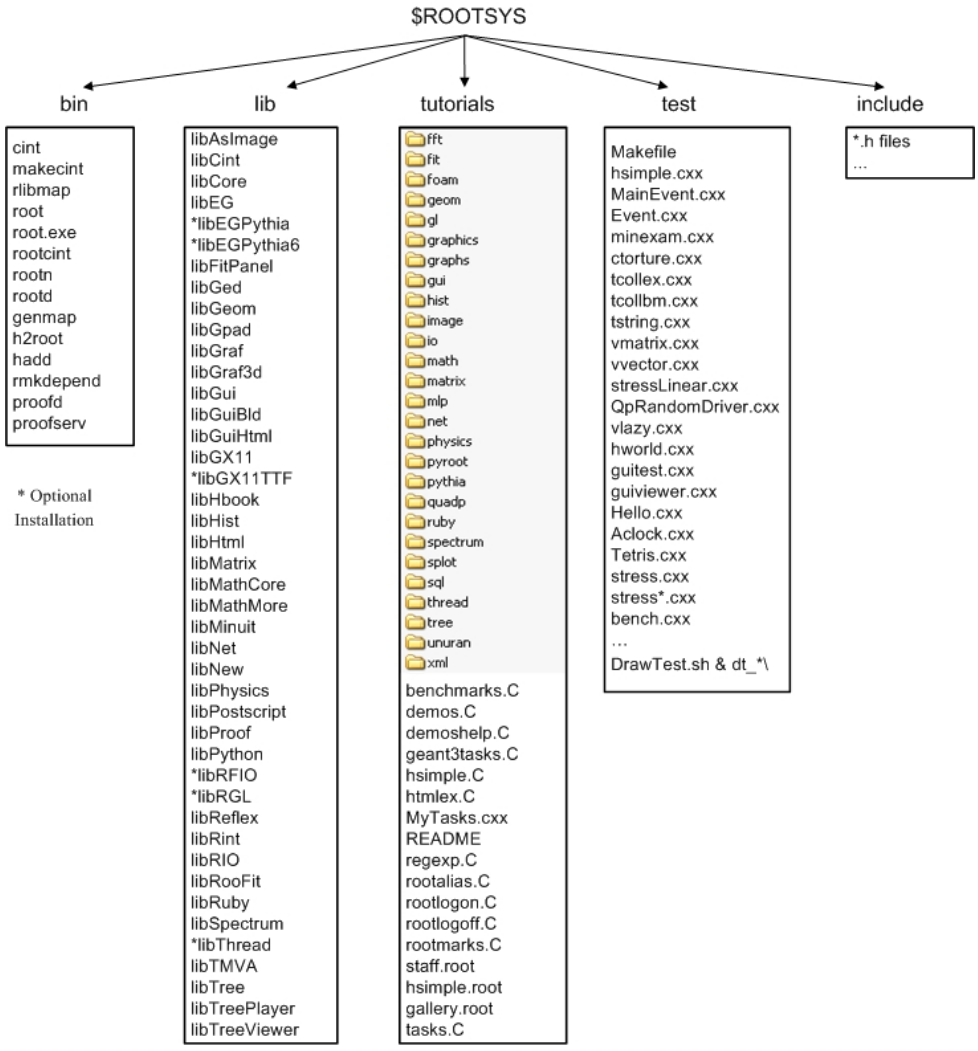

Рис. 2: ROOT frame directories

В каталоге ROOTSYS находятся примеры, исполняемые файлы, руководства, файлы руководств по заголовкам. На рисунке 2 представлено содержимое каталога. Особый интерес для меня представлял каталог tutorials. Данный каталог представляет собой с

Каталог ROOTSYS / tutorials / включает следующие подкаталоги:

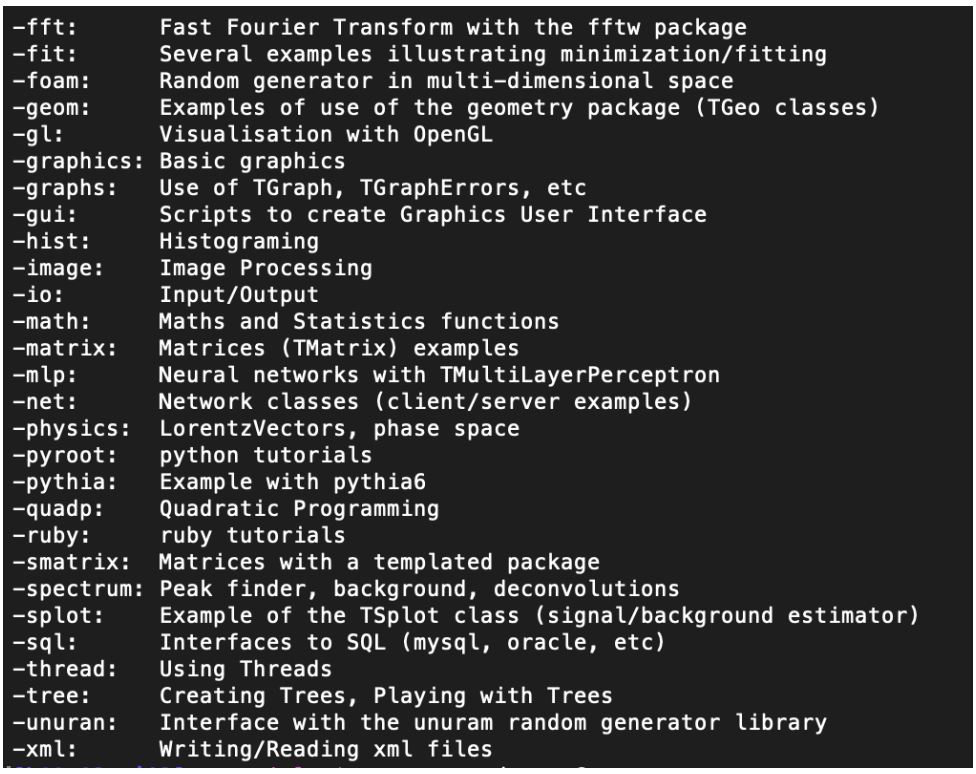

Рис. 3: The ROOTSYS/tutorials/ directory

На примере hsimple.C я ознакомилась с синтаксисом скрипта и научилась его запускать. Также ознакомилась с созданием гистограмм в ROOTe.

2 TMVA

TMVA([2]) предоставляет интегрированную в ROOT среду для обработки, параллельной оценки и применения многомерной классификации и методов многомерной регрессии. Все многомерные методы TMVA относятся к семейству алгоритмов «контролируемого обучения». Они используют обучающие события, для которых известен желаемый результат, для определения функции сопоставления, которая описывает либо границу решения, либо аппроксимацию, лежащую в основе функционального поведения, определяющую регрессию. TMVA специально разработан для анализа данных в области физики высоких энергий (HEP). В пакет входят:

- BDT;
- Линейный и нелинейный дискриминантный анализ;
- Метод максимального правдоподобия;
- SVM;
- RuleFit;
- Оптимизация прямоугольного среза, и т.д.

ROOTSYS / tutorials / tmva предоставляет примеры заданий для обучения и применения результатов обучения в классификационном или регрессионном анализе с помощью TMVA Reader.

TMVAClassification.C предоставляет пример того, как использовать обученные классификаторы в анализе. В примере использованы четыре линейно коррелированных, распределенных по Гауссу дискриминирующих входных переменных с различными средними выборками для сигнала и фона. С помощью TMVAClassification.C построены гистограммы распределения входных переменных.

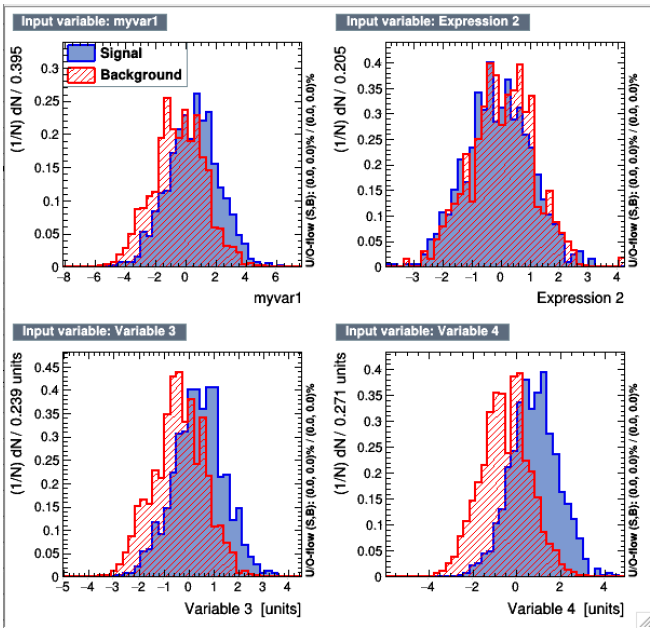

Рис. 4: Гистограммы распределения входных переменных

3 Методы классификации

Изучение различных методов классификации обусловлено тем, что нет наилучшего метода. Если алгоритм хорошо справляется с некоторыми проблемами, то он платит за это другими задачами. (Теорема о бесплатных завтраках)

3.1 Метод максимального правдоподобия

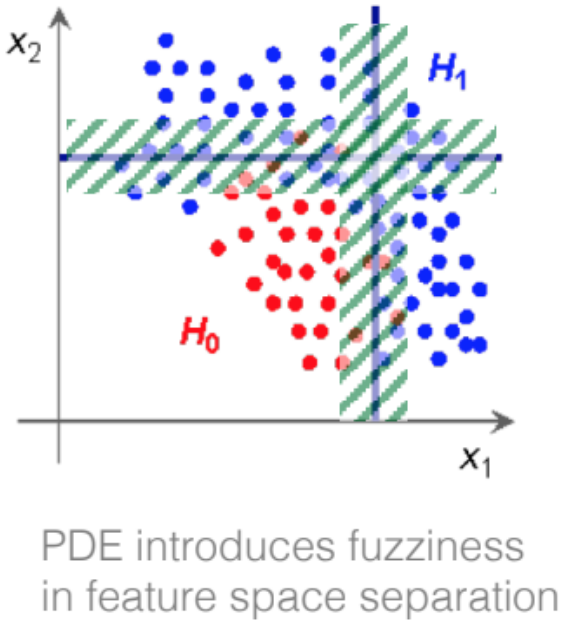

Рис. 5:

Метод максимального правдоподобия состоит в построении модели из функций плотности вероятности(PDF), которая воспроизводит входные переменные для сигнала и фона. Для данного события вероятность того, что оно относится к типу сигнала, получается путем умножения плотностей вероятностей сигнала всех входных переменных, которые считаются независимыми, и нормирования их на сумму вероятностей сигнала и фона.
Отношение правдоподобия $y_{\mathcal{L}}(i)$ для события i определяется выражением:

$$y_{\mathcal{L}}(i) = \frac{\mathcal{L}_S(i)}{\mathcal{L}_S(i) + \mathcal{L}_B(i)}, \tag{1}$$

где

$$\mathcal{L}_{S(B)}(i) = \prod_{k=1}^{n_{var}} p_{S(B),k}(x_k(i)). \tag{2}$$

Причем PDF удовлетворяет условиям нормировки:

$$\int_{-\infty}^{+\infty} p_{S(B),k}(x_k)dx_k = 1, \forall k. \tag{3}$$

Данный метод игнорирует корреляцию между входными переменными. Может использовать:

- параметрическую подгонку по функциям
- непараметрическую подгонку
- подсчет событий / гистограмма

С помощью TMVAClassification.C был построен график зависимости эффективности от разреза($N_S = N_B = 1000$):

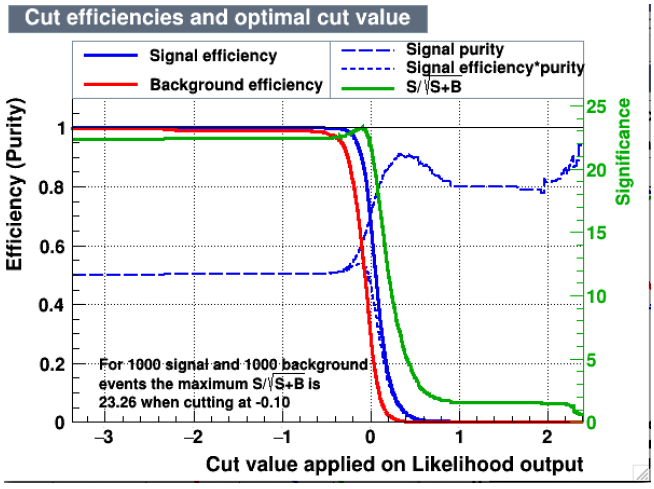

Рис. 6: Эффективность отбора и значение оптимального разреза

3.2 Линейный дискриминантный анализ Фишера

В методе дискриминантов Фишера отбор событий выполняется в преобразованном пространстве переменных с нулевыми линейными корреляциями путем различения средних значений распределений сигнала и фона. Линейный дискриминантный анализ определяет ось в (коррелированном) гиперпространстве входных переменных, так что при проецировании выходных классов (сигнал и фон) на эту ось они отодвигаются как можно дальше друг от друга, в то время как события одного и того же класса содержатся в непосредственной близости. Свойство линейности этого классификатора отражается в метрике, с помощью которой определяется ковариационная матрица пространства дискриминирующих переменных.

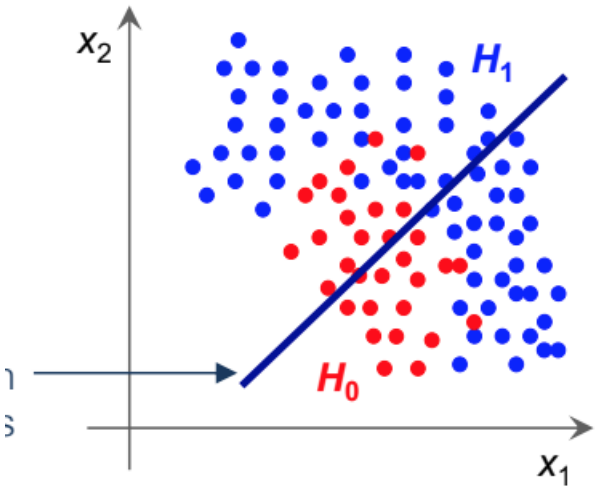

Рис. 7:

Коэффициенты Фишера, F_k, тогда даются как

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{\ell=1}^{n_{var}} W_{k\ell}^{-1} (\bar{x}_{S,\ell} - \bar{x}_{B,\ell}), \tag{4}$$

где $N_{S(B)}$ число сигнальных(фоновых) событий, $W_{k\ell}^{-1}$ - внутриклассовая матрица. Дискриминант Фишера

задается формулой

$$y_{Fi}(i) = F_0 + \sum_{k=1}^{n_{var}} F_k x_k(i). \tag{5}$$

Коэффициент F_0 - это значение среднего дискриминанта Фишера при $N_S + N_B = 0$.

С помощью TMVAClassification.C был построен график зависимости эффективности от разреза($N_S = N_B = 1000$):

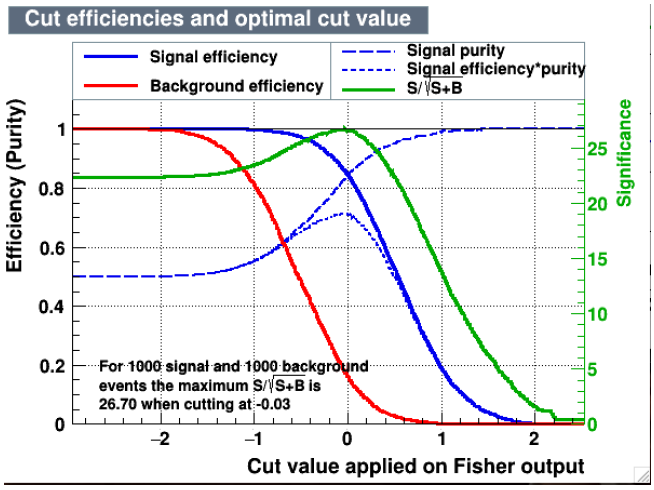

Рис. 8: Эффективность отбора и значение оптимального разреза

3.3 BDT

Дерево решений: последовательное применение разрезов разбивает данные на узлы, где последние узлы (листья) классифицируют событие как сигнал или фон большинством голосов.

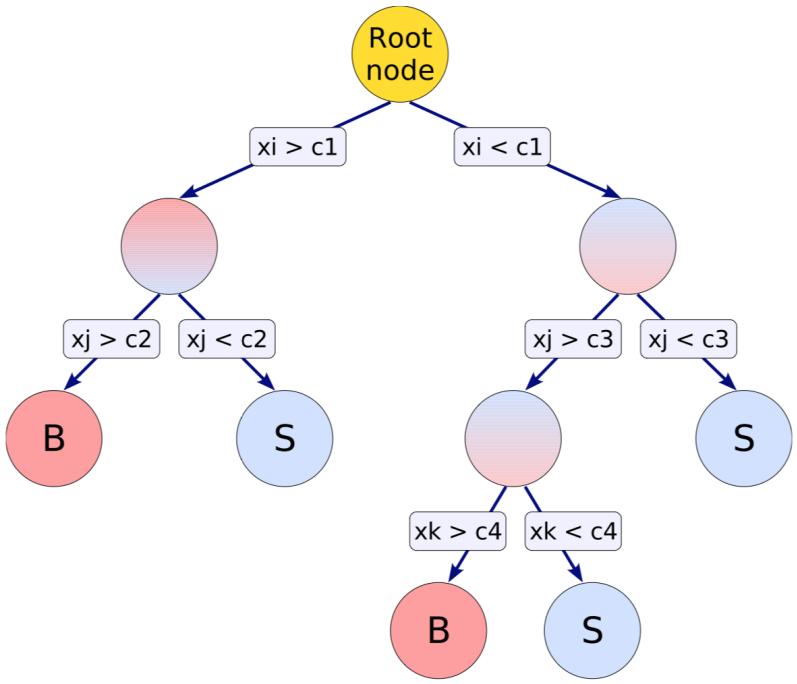

Рис. 9: Схема древовидной структуры

Построение дерева:

1. Начинается корневого узла
2. Разделяет обучающую выборку по разрезам по лучшей переменной в этом узле
3. Критерий разделения (например,$S/\sqrt{S+B}$)
4. Продолжает разделение до минимального количество событий или максимально достигнутой глубины
5. Классифицирует листовой узел по большинству событий или придает вес; неизвестное тестируемое событие классифицируются соответственно

Недостатки данного метода: нестабильность, чувствительность к перетренированности.

Критерии разделения:

- Gini Index определяется $p(1-p)$;
- Перекрестная энтропия определяется $-p\ln p-(1-p)*\ln 1-p$;
- Статистическая значимость определяется $S/\sqrt{S+B}$

Усиленные деревья решений (1996): объедение множеств деревьев решений в лес с разными взвешенными событиями в каждом дереве (сами деревья также могут быть взвешены). Главной идеей данного метода является выделение различных особенностей в выборке данных(например, трудно классифицированные события).

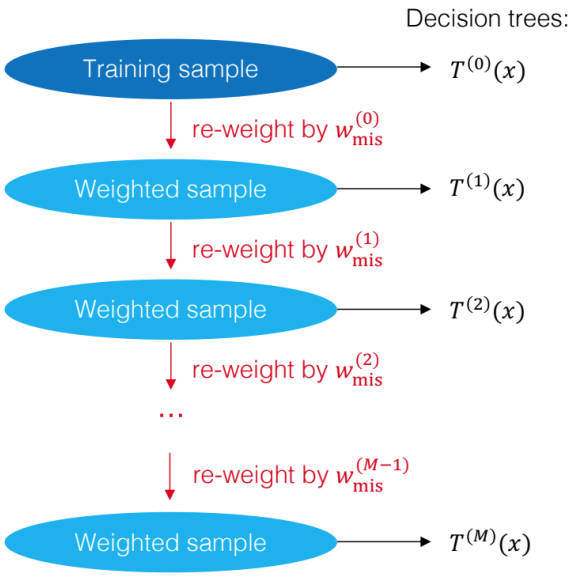

Рис. 10: Boosting

AdaBoost повторно взвешивает события, неправильно классифицированные предыдущим классификатором, по:

$$w_{mis}^{(i)} = \frac{1 - f_{mis}^{(i)}}{f_{mis}^{(i)}} \tag{6}$$

где

$$f_{mis}^{(i)} = \frac{\text{No. of misclassified events}}{\text{No. of all events}} \tag{7}$$

Окончательный BDT получен из (взвешенного) суммы по всем деревьям решений:

$$y(x) = \sum_{i=1}^{M} \ln w_{mis}^{(i)} * T^{(i)}(x) \tag{8}$$

С помощью TMVAClassification.C был построен график зависимости эффективности от разреза ($N_S = N_B = 1000$):

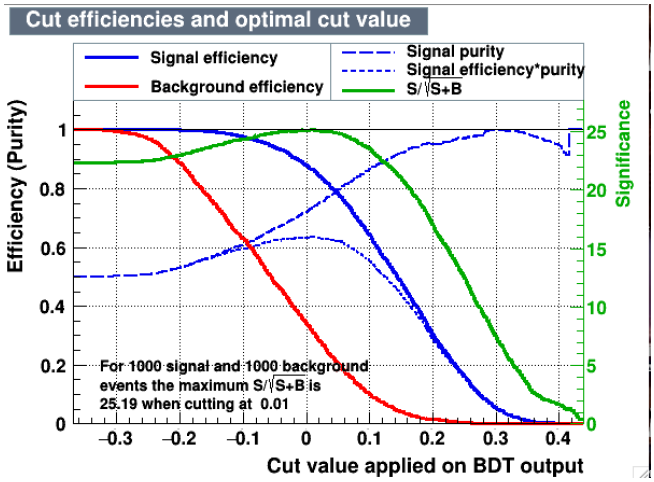

Рис. 11: Эффективность отбора и значение оптимального разреза

4 ROC - кривые

Кривая ошибок или ROC-кривая – графичекая характеристика качества бинарного классификатора, зависимость доли верных положительных классификаций от доли ложных положительных классификаций при варьировании порога решающего правила.

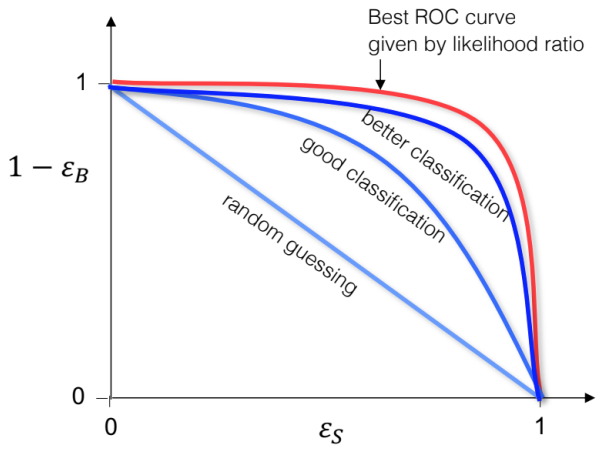

Рис. 12: Roc-кривые

Чем выше кривая ошибок, тем лучше качество классификации. С помощью TMVAClassification.C был построен график Roc-кривых для методов BDT, Fisher и Likelihood. На его основе можно сказать, что кривая ошибок метода Fisher расположена выше, чем кривые ошибок остальных методов. Значит, лучшее качество классификации в данном случае у метода Fisher.

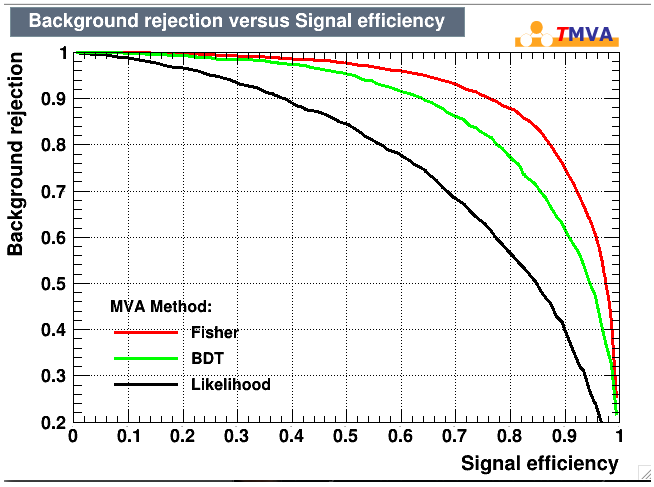

Рис. 13: Roc-кривые

Заключение

В данной работе были изучены основы программного пакета ROOT и методы многомерного анализа: BDT, Fisher и Likelihood. Построены графики Roc-кривых для различных методов классификации. На основе этих кривых сделан вывод о методе с наилучшим качеством классификации для данной задачи. В дальнейшем планируется применение методов классификации к практическим задачам.

Список литературы

[1] Официальный сайт ROOT: https://root.cern.

[2] Andreas Hocker et al. TMVA - Toolkit for Multivariate Data Analysis. 3 2007.