Tutorial on Multivariate Methods (TMVA)

iSTEP 2014
IHEP, Beijing
August 20-29, 2014

Glen Cowan (谷林·科恩)
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Introduction

Exercise: use samples of “toy” signal/background MC training data to train a multivariate classifier using the program TMVA.

Problem sheet on:

```
www.pp.rhul.ac.uk/~cowan/stat/beijing14/
istep_tmva_problem.pdf
```

Code for exercise:

```
www.pp.rhul.ac.uk/~cowan/stat/root/tmva/
tmvaExamples.tar
```

Download to local directory and unpack:

```
tar -xvf tmvaExamples.tar
```

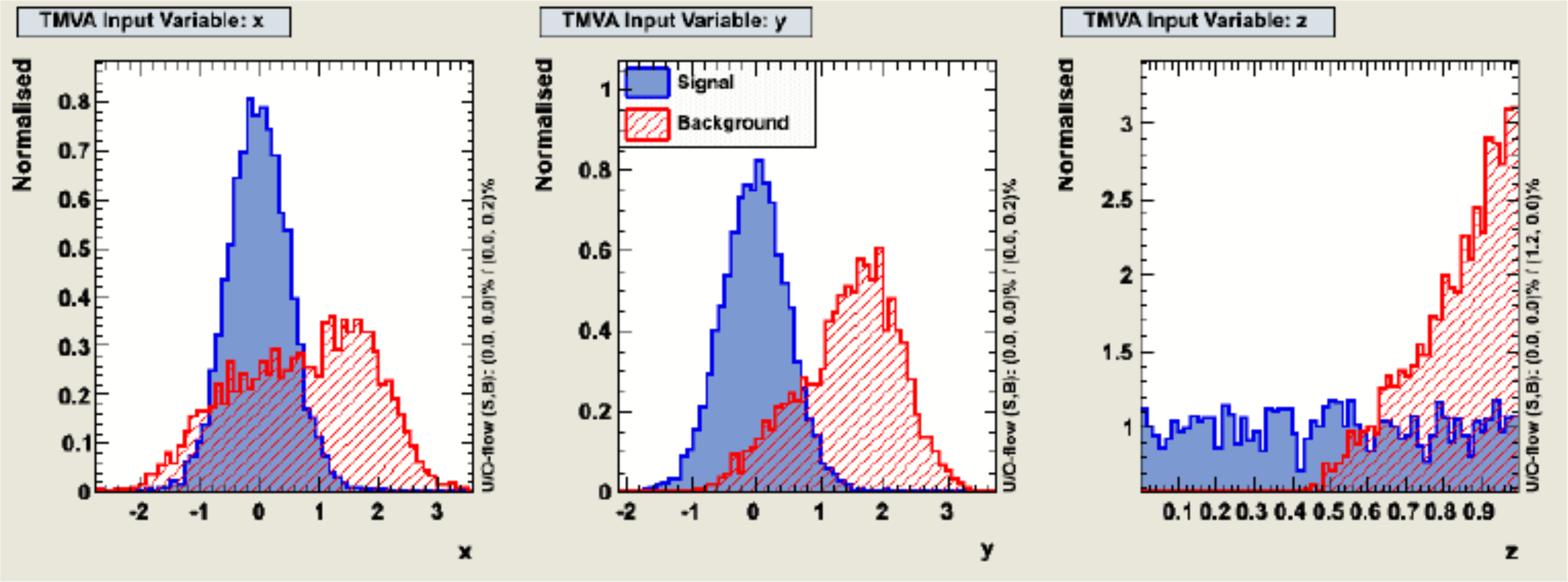
Setup for Root/TMVA at IHEP:

```
source /home/atlas01/bin/root/bin/thisroot.sh
```

Read the file `readme.txt` for more instructions.

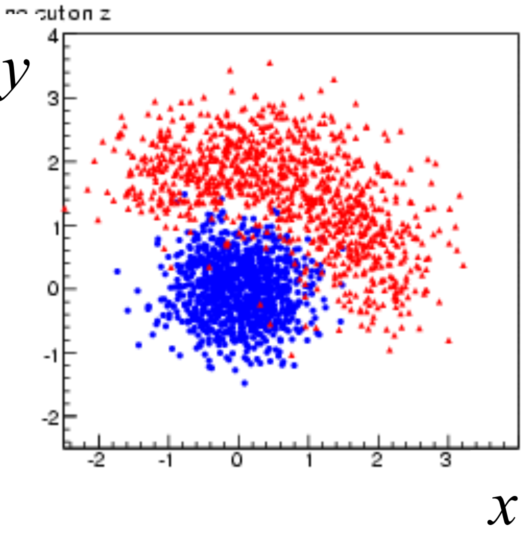
Test example with TMVA

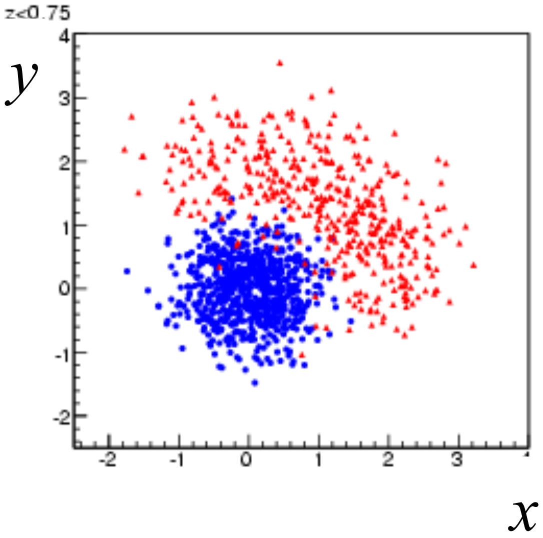
Each event characterized by 3 variables, x, y, z:



Test example (x, y, z)

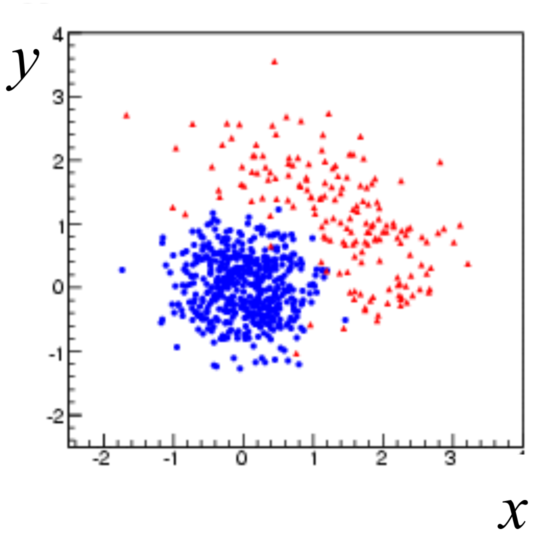
no cut on z

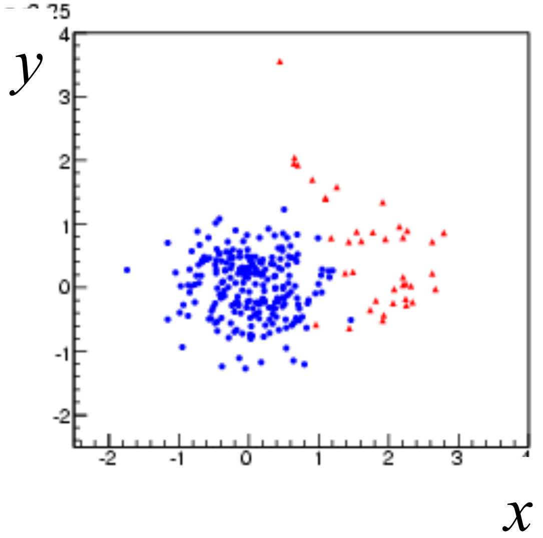




$z < 0.75$

$z < 0.5$

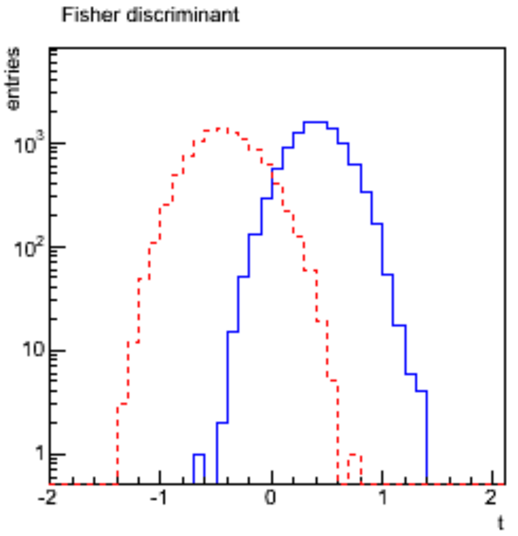


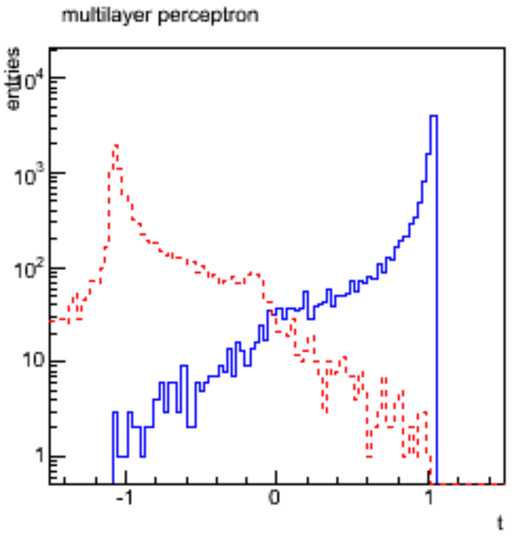


$z < 0.25$

Test example results

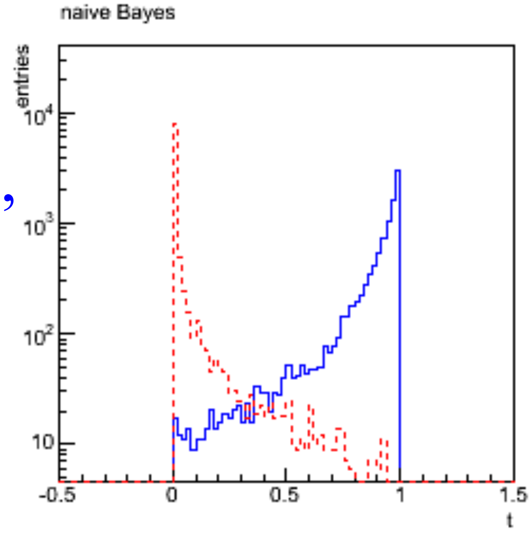
Fisher discriminant

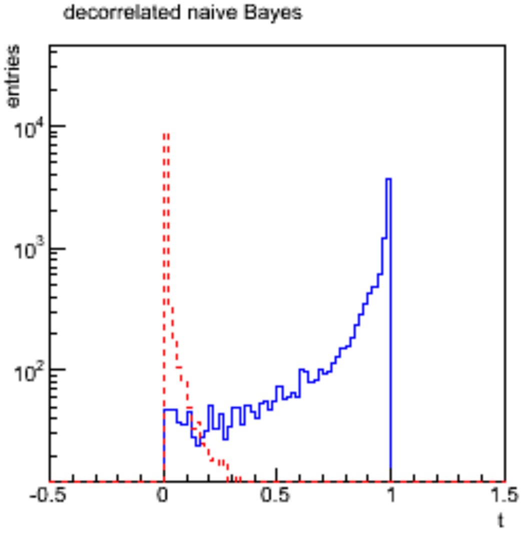




Multilayer perceptron

Naive Bayes, no decor-relation

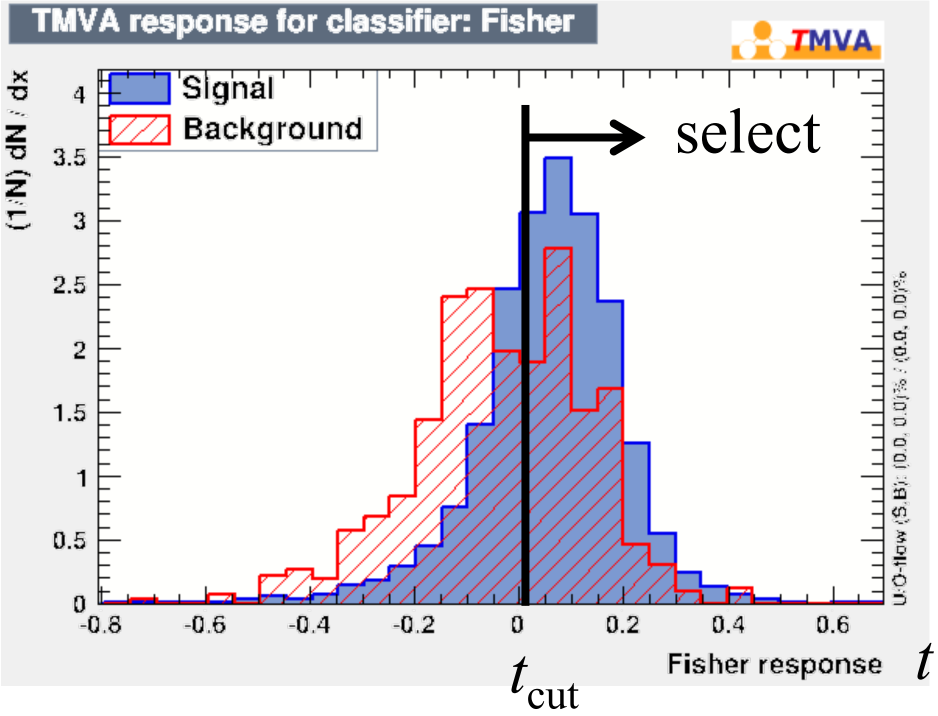




Naive Bayes with decor-relation

Extension of ATLAS Project

For the ATLAS Group Project, you defined a test statistic t to separate between signal (ttH) and background events.



You selected events with $t > t_{\text{cut}}$, calculated s and b, and estimated the expected discovery significance using $s/\sqrt{b}$.

This is OK for a start, but does not use all of the available information from each events value of the statistic t.

Likelihood ratio statistic for discovery test

In bin *i* of test statistic *t*, expected numbers of signal/background:

$$s_i = s_{\rm tot} P(t \in \text{ bin } i | s) \qquad b_i = b_{\rm tot} P(t \in \text{ bin } i | b)$$

Likelihood function for strength parameter μ with data $n_1, \ldots, n_N$

$$L(\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}$$

Statistic for test of $\mu = 0$:

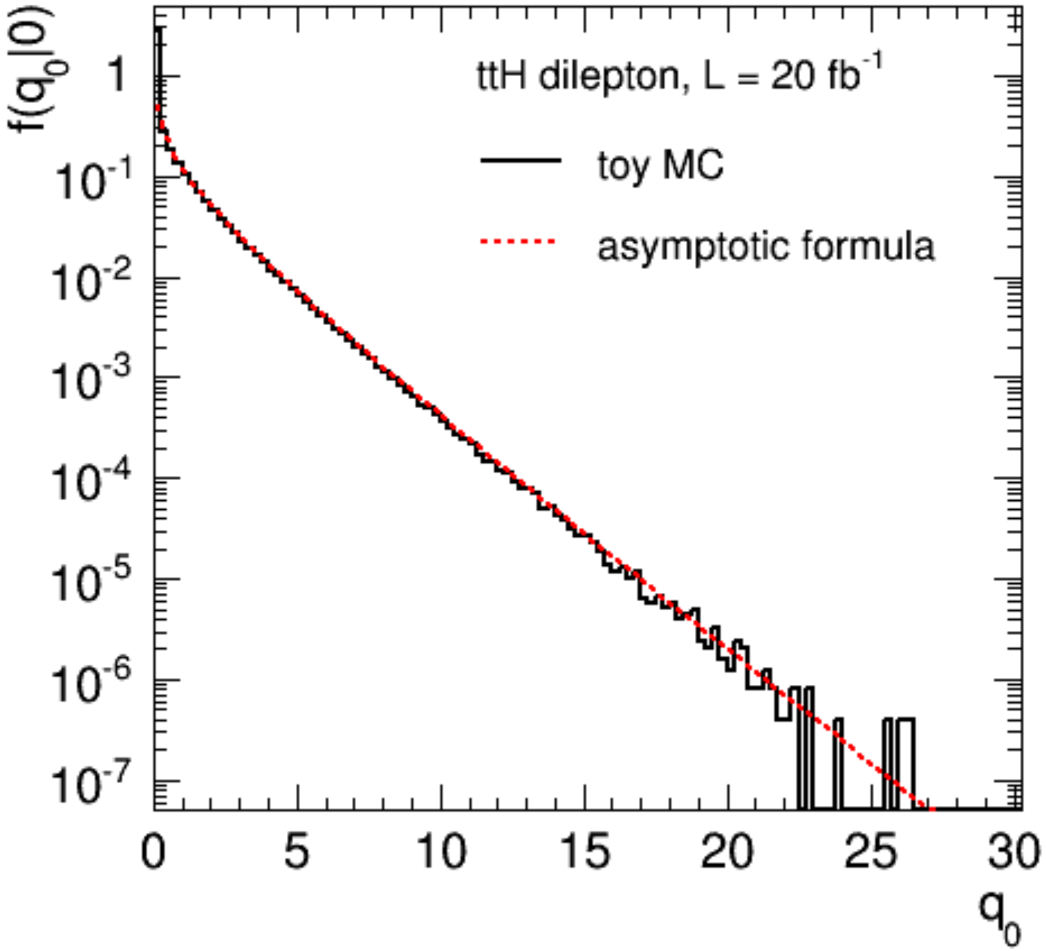
$$q_0 = \begin{cases} -2 \ln (L(0)/L(\hat{\mu})) & \hat{\mu} \geq 0 , \\ 0 & \text{otherwise.} \end{cases}$$

(Asimov Paper: CCGV EPJC 71 (2011) 1554; arXiv:1007.1727)

Background-only distribution of q_0

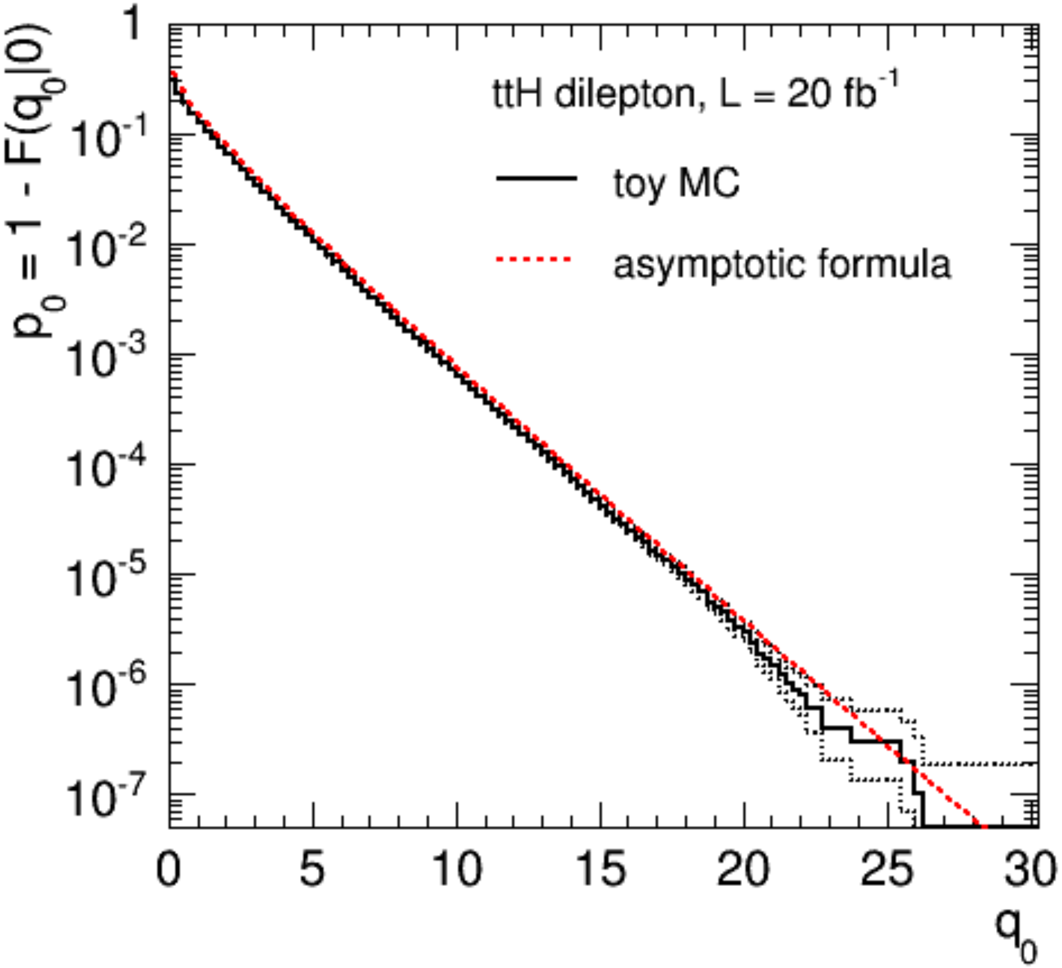
For background-only ($\mu = 0$) toy MC, generate $n_i \sim \text{Poisson}(b_i)$.

Large-sample asymptotic formula is “half-chi-square”.



Background-only cumulative distribution of q_0

p-value is probability, assuming $\mu = 0$, to find q_0 even higher than the one observed (one minus cumulative distribution).



From p-value, equivalent significance:

$$Z = \Phi^{-1}(1-p)$$

Φ^{-1} = standard normal quantile

Discovery sensitivity

Good agreement between toy MC and large-sample formulae, so OK to use asymptotic formula for significance Z,

$$Z = \sqrt{q_0}$$

Median significance of test of background-only hypothesis under assumption of signal+background from "Asimov data set":

$$n_i \to s_i + b_i$$

You can use the Asimov data set to evaluate q_0 and use this with the formula $Z = \sqrt{q_0}$ to estimate the median discovery significance.

Multinomial model

Nominal ATLAS analysis uses the information from the distribution of the MVA ouput ("shape information").

No info is taken from the total observed number of events (presumably because the systematic uncertainty on *b* is large).

This corresponds to using a multinomial model for the observed histogram of MVA output values:

$$L(\mu) = \frac{n!}{n_1! n_2! \cdots n_N!} \prod_{i=1}^{N} p_i^{n_i}$$

$$p_i = \frac{\mu s_i + b_i}{\mu s_{\rm tot} + b_{\rm tot}}$$