Searching a needle in the haystack: multivariate analysis of HEP data

Andrzej Zemla

Institute of Nuclear Physics PAS

Krakow, Poland

L Cracow School of Theoretical Physics

09-19 VI 2010 Zakopane

Motivation

- Today's experiments are bigger and much more complicated
  - We are looking for rarly proccesses
  - produce huge amount of data (mostly overhelming background)
  - Events have more discriminating variables
- Currently used alghorithms as: cuts, likelihood, Neural Networks, are well understood and easy to apply, but except NN there are not very efficient (but necessary for early data analysis)

Conclusion

Today's HEP experiments requires more complex and sophisticated tools for data analysis.

***T*MVA - *Toolkit* for Multivariate Data Analysis with ROOT**

http://tmva.sf.net

Great package which conteins most known and used alghoritms: Cuts, Likelihood, NNs, **PDE-RS, BDT, Genetic Algorithms, SVM**

Introduction

Multivariate methods can be used for:

- Classification (select signal out of background)
- Function approximation (regression)
- Probability density estimation (estimate the probability distribution)
- Variable selection (find most important variables)
- Optimization (tuning, verification)
- Many others...

In our group are using them for tau identification in ATLAS – we are doing classification.

Bayesian vs. Frequentist approach

- **PROBABILITY: degree of belief** (Bayes, Laplace, Gauss, Jeffreys, de Finetti)
- **PROBABILITY: relative frequency** (Venn, Fisher, Neyman, von Mises).
- **Bayesian approach:** probability is degree of belief. Thus the probability *p* is our assessment of the probability of success at each trial, based on our current state of knowledge.

  If our assessment, initially, is incorrect? As our state of knowledge changes, our assessment of the probability of success changes accordingly.
- **Bayesian inference** is statistical inference in which **evidence or observations are used** to update or to newly infer the probability that a hypothesis may be true.
- This allows for a *cleaner* foundation than the frequentist interpretation.

"*We don't know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some a priori postulates. My problem is to get these stated as clearly as possible.*"

Sir Harold Jeffreys, in a letter to Sir Ronald Fisher dated 1 March, 1934

H.B. Prosper,"Bayesian Analysis", arXiv:hep-ph/0006356v1 30 Jun 2000

Bayes Theorem

- Bayes' theorem relates the conditional (posterior) and marginal (prior) probabilities of events A and B:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- **P(A)** is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

- Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.

Multivariate Methods

Machine Learning

Teach a machine to learn $y = f(x)$ by feeding it **training data** $\boldsymbol{T} = (\boldsymbol{x}, \boldsymbol{y}) = (x,y)_1, (x,y)_2,\ldots(x,y)_N$ and a **constraint** on the class of functions.

Bayesian Learning

For each function f(x) in the function space F calculate the posterior probability $p(\boldsymbol{f} | \boldsymbol{T})$ using a given training sample T= $(\boldsymbol{x}, \boldsymbol{y})$.

Don't use a single function, use a bunch of functions weighted by probabilities

- The posterior probability is the conditional probability that is assigned after the relevant evidence is taken into account.
- Training sample: $\boldsymbol{T} = (\boldsymbol{x}, \boldsymbol{y})$ set of input vectors $\boldsymbol{x}$ and desired outputs $\boldsymbol{y}$.

Bayesian Learning: Why?

- **Probabilistic learning:** Calculate explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems.
- **Probabilistic prediction**: Predict multiple hypotheses, weighted by their probabilities.
- **Incremental:** Each training example can incrementally increase or decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.
- **Standard:** Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

Classification

A Bayes classifier:

$$p(S|x) = \frac{p(x|S)\,p(S)}{p(x|S)\,p(S) + p(x|B)\,p(B)}$$

where ***S*** is associated with *y* = **1** and ***B*** with *y* = **0**. **Bayes classifier** *accepts events x if p(**S**|x) > **cut** as belonging to **S**.*

We need to approximate probability distributions *P(**x**|**S**)* and *P(**x**|**B**)*.

- *If your goal is to **classify objects** with the fewest errors, then the **Bayes classifier** is the **optimal** solution.*
- Consequently, if you have a classifier known to be ***close*** to the **Bayes limit**, then *any* other classifier, *however sophisticated*, can ***at best*** be only marginally better than the one you have.
  - =>If your problem is **linear** you don't gain anything by using sophisticated **Neural Network**
- *All* classification methods, such as the ones in TMVA, are different numerical approximations of the Bayes classifier.

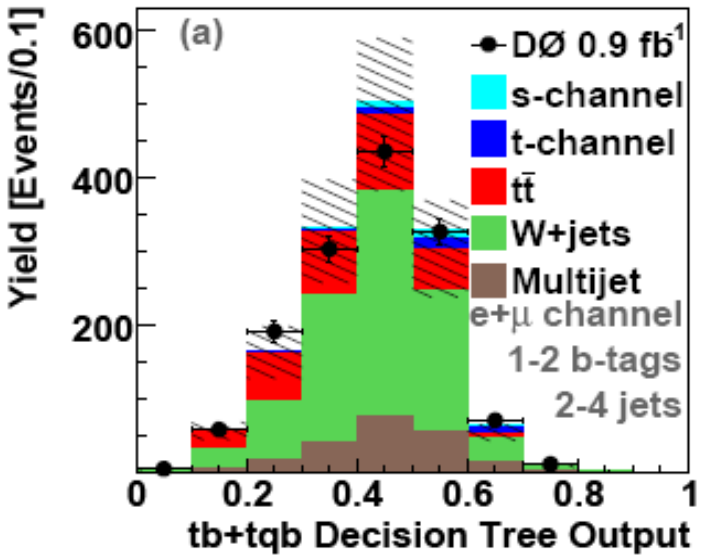
Practical applications

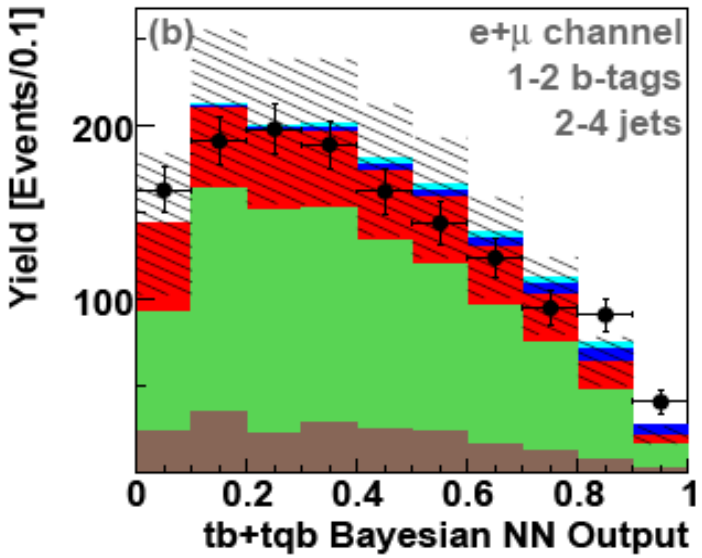
A Short List of Multivariate Methods

- Cuts
- Linear Discriminants (like Fisher)
- Quadratic Discriminants
- **Support Vector Machines**
- Naive Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- **Decision Trees**
- Neural Networks
- **Bayesian Neural Networks**
- **Genetic Algorithms**

- And many, many others.....

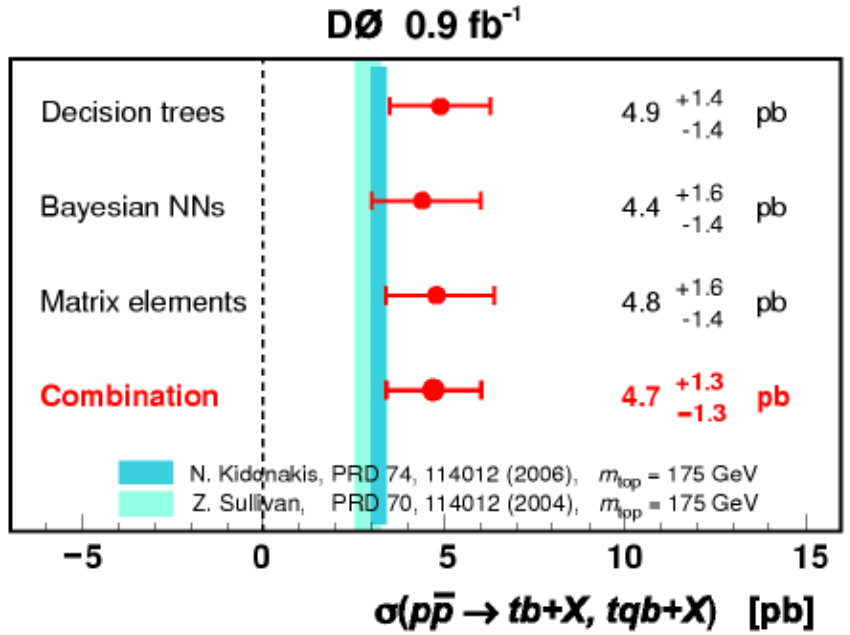
Example – Single Top Search





Single top quark search using
boosted decision trees
Bayesian neural networks

D0 Collaboration,
PRD 78 012005, 2008



Discriminant Verification

PROBLEM: How can one confirm that the *n-dimensional* density is well-modeled?

Any classifier ***f(x)*** close to the Bayes limit approximates

$D(x) = p(x|S) / [\, p(x|S) + p(x|B) \,]$ (Bayes discriminant)

Therefore, if we weight, ***event-by-event***, an admixture of N signal and N background events by the function ***f(x)***

$S_w(x) = N\, p(x|S)\, \boldsymbol{f(x)}$

$B_w(x) = N\, p(x|B)\, \boldsymbol{f(x)}$

then the sum

$S_w(x) + B_w(x) = N\,(p(x|S) + p(x|B))\, \boldsymbol{f(x)} = \boldsymbol{N\, p(x|S)}$, i.e., we should recover the n-dimensional ***signal density***.

A check, based on Monte Carlo, that the probability density is well modeled.

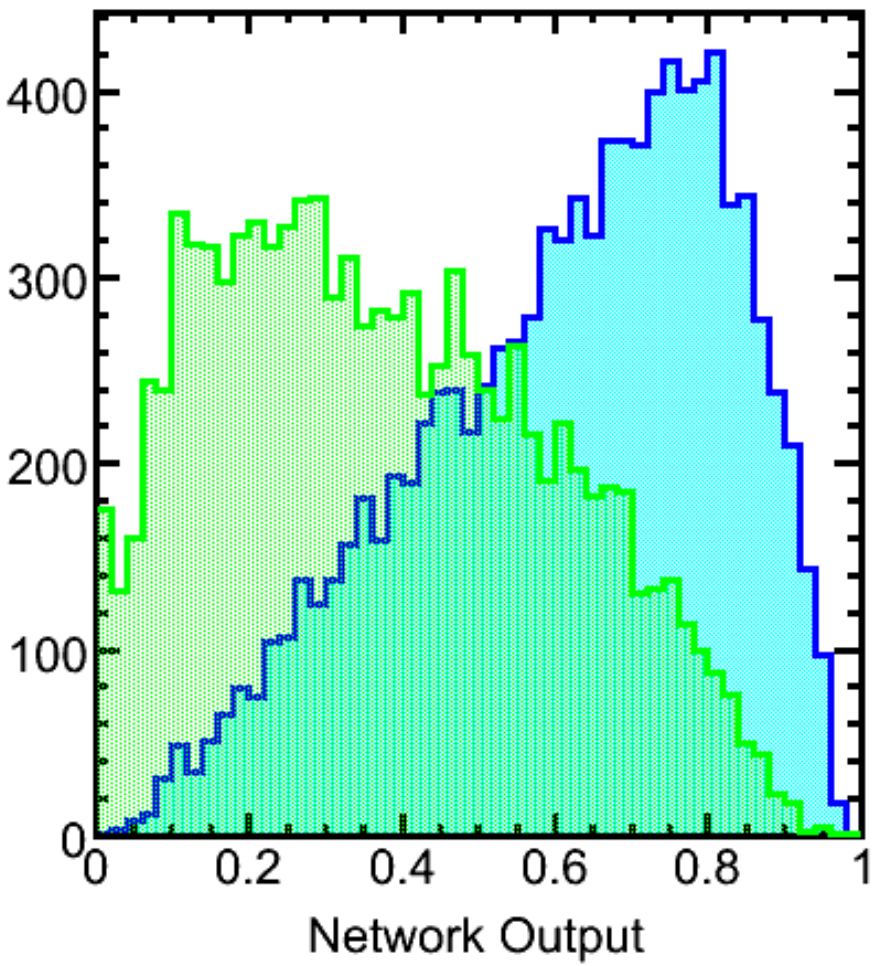
Verification – Example

Dzero single top quark search

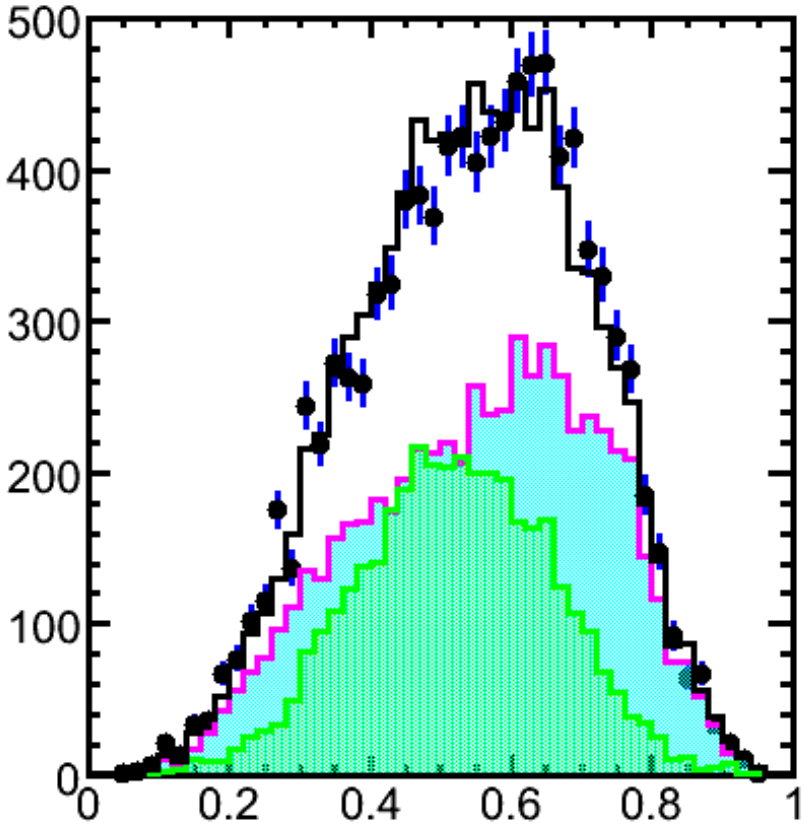
Verifying the Bayesian neural network discriminant.

Number of input variables ~ 24
Number of channels = 12

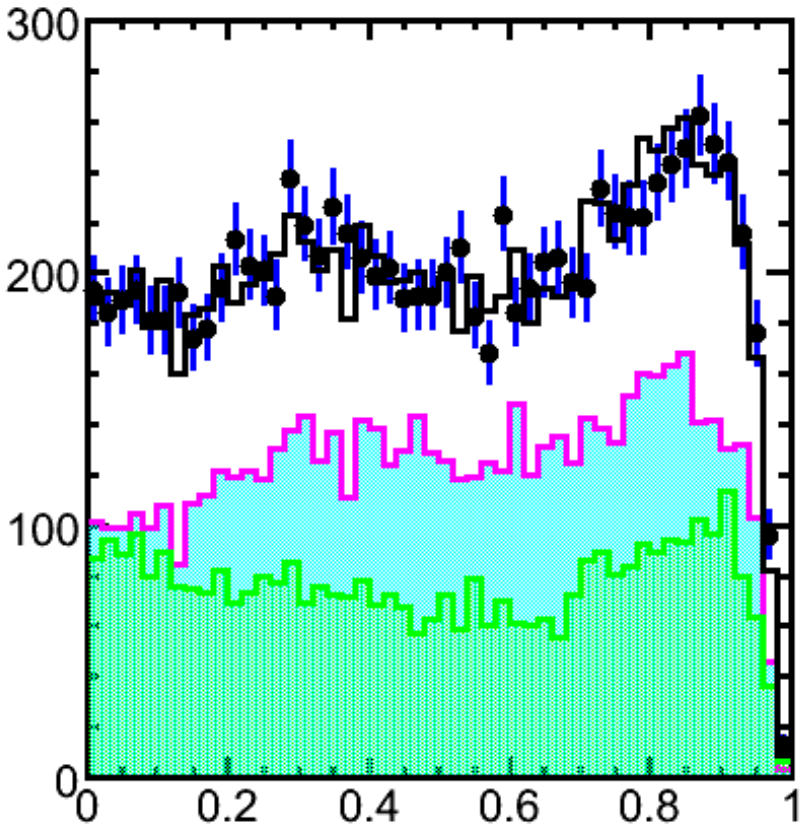
(*e*, μ) x (1, 2) b-tags x (2,3,4) jets



Verification – Example



$Q_{lepton} \times \eta_{leading\ untagged\ jet}$



$\cos(\text{lepton}, Q \times Z)_{best\ top}$

Cyan plot: weighted signal
Green plot: weighted background
Black curve: sum
Black dots: signal

Systematics – some concepts

Same steps as using cut based analysis:

Systematic errors:

- Vary all aspects of the model about which one is uncertain and run the analysis.
- Get distribution of results (ex. Higgs masses)
- Usual error propagation fails – non-linearity of the error propagation

Robustness (stability):

- Different or differently trained MV leads in general to different results.
- Solution should be unaffected by these variations.

"In summary, handling systematic errors with MV-based analysis is no different from the procedure for cut-based analysis 99% of the effort, as you know, is convincing yourself that your N-dimensional model agrees sufficiently well with the N-dimensional data to be trustworthy!"

H. B. Prosper

Summary

- Multivariate methods can be applied to many aspects of data analysis.
- Many practical methods, and convenient tools such as TMVA, are available for regression and classification.

- **All methods approximate the same mathematical entities, but no one method is guaranteed to be the best in all circumstances. Simplicity, speed of learning and robustness also matters. So, experiment with a few of them!**

- **We need methods, and convenient tools, to explore and quantify the quality of modeling of n-dimensional data.**

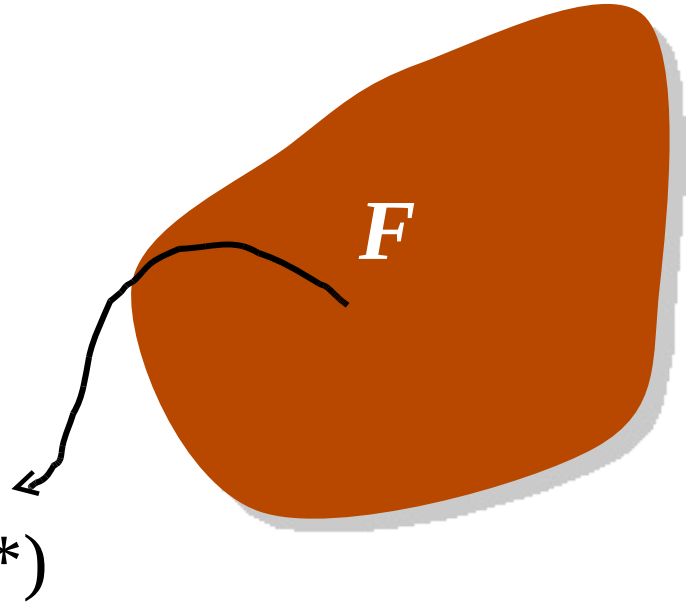
More general question

- **Is there a sensible way to use multivariate methods when one does not know for certain where to look for signals?**

Machine Learning

We have to chose:

Function class	***F* = { *f*(*x*, *w*) }**
Constraint	***C***
Loss function	***L***
	example:
	$L=y_i^2 - f^2(x_i,w)$
Training data	***T*(*y*,*x*)**



Method

Find *f*(*x*, ***w***) by minimizing the **empirical risk *R***

$$R(w)=\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, w))$$

subject to the constraint *C*(***w***)

We choose at the end a single “best” function *f*(*x*, ***w***)
(best single Neural Network, best likelihood etc.)

Bayesian Learning

Choose

Function class	$\boldsymbol{F} = \{ f(x, \boldsymbol{w}) \}$
Prior	$p(\boldsymbol{w})$
Likelihood	$p(\boldsymbol{y}\|\boldsymbol{x}, \boldsymbol{w})$
Training data	$T\boldsymbol{(y,x)}$

Method

Use Bayes' theorem to infer the parameters:

$p(\boldsymbol{w}|\boldsymbol{T}) = \{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})\, p(\boldsymbol{x}|\boldsymbol{w})\, p(\boldsymbol{w})\}/\{p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x})\}$

because $p(w|y,x) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{x})/p(\boldsymbol{y}|\boldsymbol{x})$

$\sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) * p(\boldsymbol{w})$ (assume $p(\boldsymbol{x}|\boldsymbol{w}) = p(\boldsymbol{x})$)

~ Likelihood * Prior

We do not pick up a single function *f(x)*, instead $p(\boldsymbol{w}|\boldsymbol{T})$ assigns a probability density to every function in the function class.