



Introduzione all'Analisi Multivariata

I. MVA con TMVA

Pietro Biassoni^{1,2}

¹INFN Milano

²Dipartimento di Fisica
Università degli Studi di Milano

Metodologie di Analisi dei Dati
15 Gennaio 2010



Outline



- 1 Introduzione
- 2 Set up del vostro account
- 3 MultiVariate Analysis
- 4 Esercizio



Outline

- 1 **Introduzione**
- 2 Set up del vostro account
- 3 MultiVariate Analysis
- 4 Esercizio



TMVA

Lo scopo della lezione di oggi è di imparare ad utilizzare il pacchetto di analisi dati TMVA per effettuare una MVA.

Il campione che ci apprestiamo ad utilizzare
`/home/comune/AnalisiDati/mva2_sample.root`
contiene due alberi (*signal* e *bkg*) con otto variabili ciascuno. Le variabili hanno tutte un basso potere discriminante.

Cosa impareremo a fare:

- Inizializzare un analisi MVA con TMVA;
- Visualizzare i risultati;
- Ottimizzare l'analisi

Ricordo che esistono altri pacchetti di analisi dati molto versatili e potenti. Provate a usare (a casa) StatPatternRecognition.



Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 MultiVariate Analysis
- 4 Esercizio



Set up del vostro account

Per utilizzare l'ultima versione di TMVA dobbiamo cambiare la versione di ROOT in uso.

Nella vostra home, nel file `.zshrc` aggiungete le seguenti righe

File `~/ .zshrc`

```
export ROOTSYS=/tools/ROOTINSTALL/root_v5.26.00.Linux-slc4-gcc3.4
export LD_LIBRARY_PATH=$ROOTSYS/lib:$LD_LIBRARY_PATH
export PATH=$ROOTSYS/bin:$PATH
```

Nella vostra home, nel file `.rootrc` aggiungete la seguente riga

File `~/ .rootrc`

```
Unix.*.Root.MacroPath: $ROOTSYS/tmva/test:$ROOTSYS/macros:.
```



Controllo del vostro account

Infine controllate che tutto sia ok con i seguenti comandi

```
[biassoni@lab24 ~]% source ~/.zshrc  
[biassoni@lab24 ~]% echo $ROOTSYS  
/tools/ROOTINSTALL/root_v5.26.00.Linux-slc4-gcc3.4  
[biassoni@lab24 ~]% which root  
root='root -l'  
/tools/ROOTINSTALL/root_v5.26.00.Linux-slc4-gcc3.4/bin/root
```



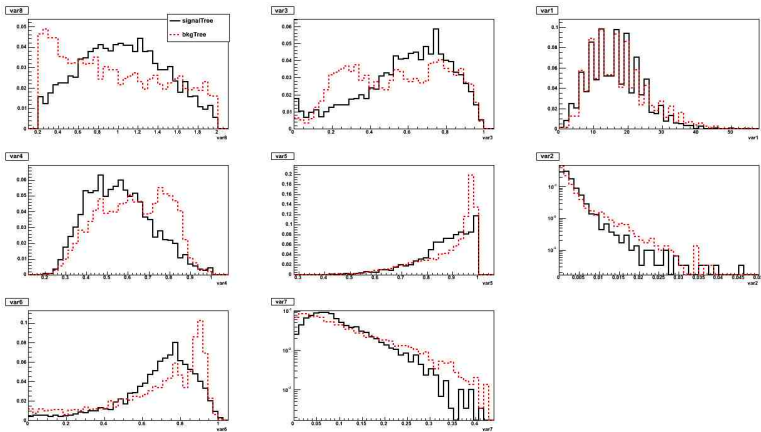
Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 MultiVariate Analysis**
- 4 Esercizio



Weak variables

Nella scorsa lezione abbiamo affrontato un caso di discriminazione segnale/fondo in presenza di variabili fortemente discriminanti. Oggi ci occuperemo di un caso in cui abbiamo molte più variabili, ma con poco potere discriminante (*weak variables*).





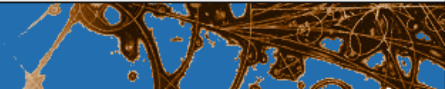
Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 **MultiVariate Analysis**
 - **TMVA Factory**
 - TMVA GUI
 - Classificatori TMVA
 - TMVA Reader
- 4 Esercizio



TMVA

Introduzione



TMVA è un pacchetto di analisi dati incluso in ROOT.

La guida del pacchetto si trova in

<http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf>

Noi copriremo una piccola parte delle potenzialità del pacchetto e impareremo a:

- Dichiarare la struttura di analisi;
- Trainare un Fisher, una NN, un BDT;
- Modificare (in parte) i parametri di questi oggetti;
- Visualizzare i risultati e confrontare le prestazioni

Per fare questo useremo una macro già pronta (scaricabile dal sito) che modificheremo nella parte relativa al training dei classificatori.



TMVA

La factory

TMVA funziona come una *black box*, la gestione dei dati e dei metodi da trainare e testare è gestita in modo completamente trasparente per l'utente dalla *factory*.

Per istanziare una *factory*

```
TMVA::Factory *factory = new TMVA::Factory( "MVAnalysis",  
outputFile, "" );
```

Per fornire i dati alla *factory*

```
factory->AddSignalTree(signalTree, signalWeight);  
factory->AddBackgroundTree(bkgTree, bkgWeight);
```

Per fornire le variabili alla *factory*

```
factory->AddVariable("var1");
```



TMVA

Booking, Training e Test dei classificatori

Una volta istanziata, la *factory* va istruita su quali classificatori volete usare

```
// Preparare i tree
factory->PrepareTrainingAndTestTree("", "nTrain_Signal=2000:
    nTrain_Background=2000:nTest_Signal=1000:nTest_Background=1000");
// Scegliere il classificatore (vedi slides successive)
factory->BookMethod( TMVA::Types::kFisher, "Fisher","");
// Trainare i classificatori
factory->TrainAllMethods();
// Testare i classificatori
factory->TestAllMethods();
// Comparare i vari classificatori
factory->EvaluateAllMethods();
```



TMVA

Ranking delle variabili

Durante il training, la *factory* manda a video molte informazioni, una delle più importanti è il ranking delle variabili a seconda del loro potere discriminante.

```

--- ANN : Ranking result (top variable is best ranked)
--- ANN : -----
--- ANN : Rank : Variable : Importance
--- ANN : -----
--- ANN : 1 : var1 : 2.309e+03
--- ANN : 2 : var3 : 2.623e+01
--- ANN : 3 : var8 : 1.818e+01
--- ANN : 4 : var5 : 1.041e+01
--- ANN : 5 : var6 : 7.615e+00
--- ANN : 6 : var4 : 4.678e+00
--- ANN : 7 : var7 : 1.089e-01
--- ANN : 8 : var2 : 1.104e-04
--- ANN : -----

```



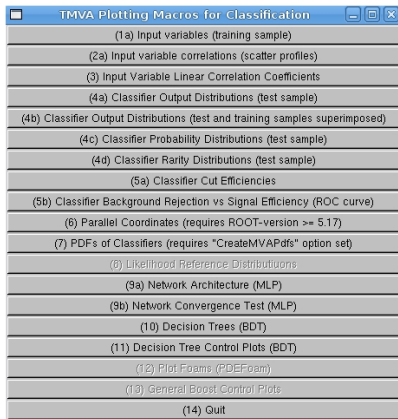
Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 **MultiVariate Analysis**
 - TMVA Factory
 - **TMVA GUI**
 - Classificatori TMVA
 - TMVA Reader
- 4 Esercizio



TMVA GUI

Dopo il training viene invocata una GUI (Graphical User Interface)



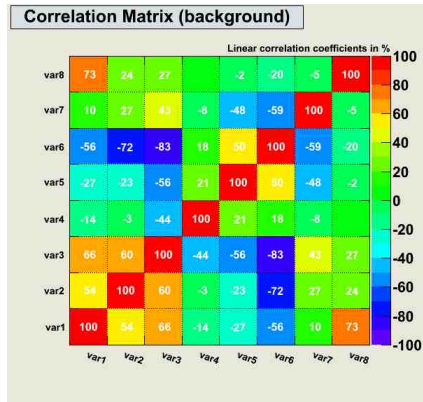
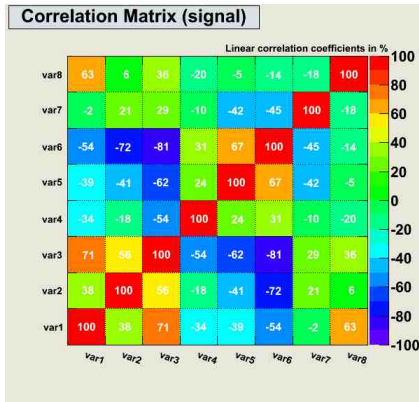


TMVA GUI

Correlazioni

In una MVA è molto importante controllare le correlazioni tra le variabili di input, infatti usare due variabili molto correlate non solo è inutile, ma potrebbe risultare dannoso.

La GUI ha un'opzione per plottare tali correlazioni.



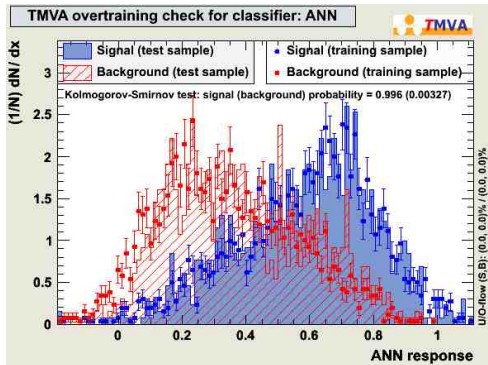


TMVA GUI

Output

Una quantità interessante è la distribuzione dell'output del classificatore, questo permette:

- di valutare *a occhio* la separazione segnale/fondo;
- di accorgersi di palesi casi di overtraining (forte discrepanza tra l'output nel campione di test e in quello di training) .

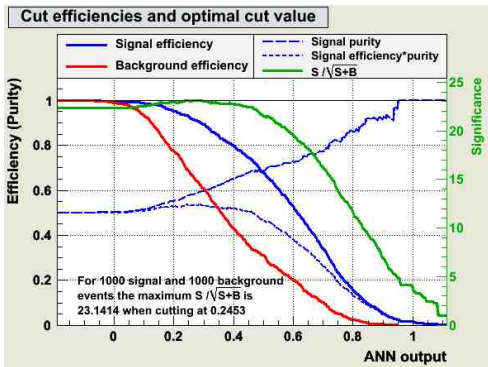




TMVA GUI

Ottimizzazione del taglio

Se voleste applicare un taglio all'output del classificatore per discriminare il segnale dal fondo, TMVA calcola per voi tutte le grandezze (efficienza segnale, reiezione fondo, ...) che abbiamo discusso nella lezione precedente



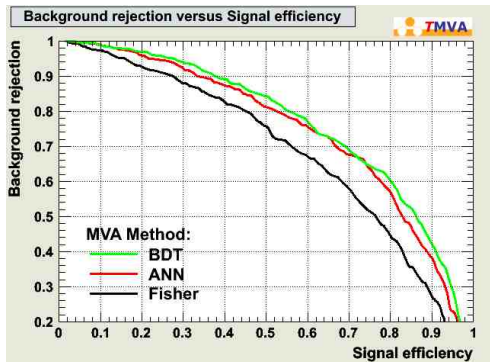


TMVA GUI

ROC curves

Lo strumento in assoluto più importante sono le ROC curves (efficienza segnale vs. reiezione del fondo).

Infatti consentono di valutare a parità di efficienza quale sia il classificatore più potente.





Outline

1 Introduzione

2 Set up del vostro account

3 MultiVariate Analysis

- TMVA Factory
- TMVA GUI
- **Classificatori TMVA**
- TMVA Reader

4 Esercizio



Classificatori TMVA

DISCLAIMER!

DISCLAIMER!!!

La MVA è un ambito in continua crescita ed espansione.

Con ogni nuova release di TMVA, i classificatori e le loro opzioni vengono migliorati o ne vengono aggiunti di nuovi. In queste slides presenterò i tre classificatori più usati e le loro opzioni principali.

Provate anche gli altri classificatori e le altre opzioni, tutto è documentato in modo estremamente intuitivo in

<http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf>



Classificatori TMVA

Fisher

Fisher

```
factory->BookMethod( TMVA::Types::kFisher, "Fisher", "");
```

Questo è il classificatore più semplice e pertanto non ha nessuna opzione.
Per scegliere questo metodo bisogna usare il comando



Classificatori TMVA

Artificial Neural Network



ANN

```
factory->BookMethod( TMVA::Types::kMLP, "ANN","...");
```

ANN Options

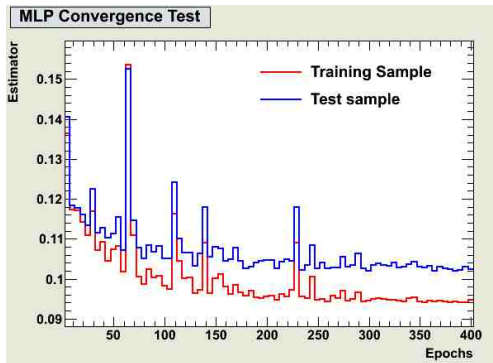
- "NCycles=400" Numero di cicli di training;
- "HiddenLayers=N+1" Struttura della rete "N+1" è un layer di N+1 neuroni, "N+1;N*N" sono due layer il primo di N+1 e il secondo di N^2 neuroni;
- "NeuronType" Funzione di attivazione del neurone;
- "NeuronInputType" Funzione di sinapsi del neurone;
- "TestRate" Numero di cicli di training tra due test;
- "DeacayRate" Parametro di decadimento.



Classificatori TMVA

Artificial Neural Network: Convergence Test

Una strumento molto utile per valutare il possibile overtraining della rete è la differenza tra l'errore del classificatore sul campione di training e su quello di test in funzione del numero di cicli di training.



Dopo le fluttuazioni iniziali, la differenza deve restare costante.



Classificatori TMVA

Boosted Decision Tree

BDT

```
factory->BookMethod( TMVA::Types::kBDT, "BDT", "" )
```

BDT Options

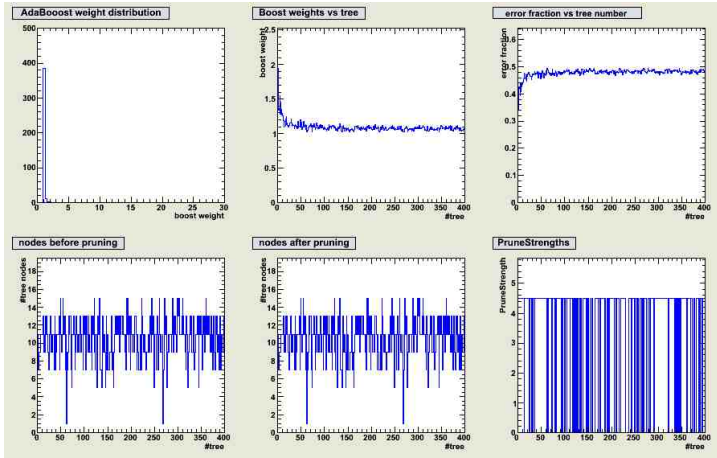
- "nTrees" Numero di alberi;
- "BoostType" Algoritmo di Boosting;
- "SeparationType" FOM che viene ottimizzata in ogni nodo;
- "nEventsMin" Numero di eventi minimo in una foglia;
- "nCuts" Passo dello scan sui tagli per ogni variabile e ogni nodo;
- "UseYesNoLeaf" Output logico (1/0) o analogico (peso);
- "UseWeightedTree" Output pesato o voto di maggioranza;
- "PruneMethod" Criterio di Pruning;
- "PruneStrenght" Prune parameter.



Classificatori TMVA

Boosted Decision Tree Convergence Test

Alcuni test standard vengono effettuati sui BDT, la regola generale è che le distribuzioni devono essere asintoticamente piatte per evitare l'overtraining.





Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 **MultiVariate Analysis**
 - TMVA Factory
 - TMVA GUI
 - Classificatori TMVA
 - **TMVA Reader**
- 4 Esercizio



TMVA Reader

Una volta completato il training, i pesi utilizzati internamente da TMVA vengono salvati nella directory *weights*.

Per calcolare il valore del classificatore su un dataset diverso da quello di training e di testing, si utilizza TMVA Reader

TMVA Reader

```
// Istanzia il reader
TMVA::Reader *reader = new TMVA::Reader("");
// Aggiunge le variabili usate dal classificatore
reader->AddVariable("varname_used_during_training",&variable);
// Sceglie il classificatore
reader->BookMVA("MVAname_used_during_training",
               "MVA_weight_file.xml");
// Ottiene il valore
double value =
reader->EvaluateMVA("MVAname_used_during_training")
```



Outline

- 1 Introduzione
- 2 Set up del vostro account
- 3 MultiVariate Analysis
- 4 **Esercizio**



Esercizio



Questo esercizio è più semplice e “divertente” da fare rispetto agli scorsi, purtroppo è più difficile documentare il lavoro fatto...



Come svolgere una MVA

La cattiva notizia è che non c'è mai una MVA giusta e una sbagliata. Ci sono di sicuro cose da fare e da non fare, ma mai una garanzia sul risultato.

Cose da fare

- Plottare le variabili
- Plottare le correlazioni
- Provare vari classificatori con vari parametri
- Provare a togliere/aggiungere le variabili correlate
- Provare a togliere/aggiungere le variabili meno discriminanti
- Confrontare i classificatori
- Controllare l'overtraining i classificatori



Come svolgere l'esercizio

L'esercizio consiste nel *giocare* con TMVA.

- Provate i metodi spiegati (facoltativi quelli non spiegati), cercando di ottenere la massima resa dal classificatore.
- Studiate almeno un caso di overtraining della ANN e del BDT.
- Documentate quello che fate.

Come documentare? A voi la scelta tra mail chilometrica con figure incluse (sconsigliato un .tar con dentro 50 figure....) oppure un (piccolo) file .pdf con testo che spiega cosa avete fatto e le figure.

Il mio esercizio è giusto? Non c'è una soluzione giusta o sbagliata! La parte importante è provare, provare provare! E giustificare i risultati.

Vorrei pubblicare le soluzioni più interessanti sul sito (prego scrivermi nella vostra mail il consenso per farlo e se autorizzate uso di solo nome solo cognome o entrambi).