View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Almae Matris Studiorum Campus

ALMA MATER STUDIORUM - Università di Bologna

69703 - ANALISI STATISTICA DEI DATI NELLA FISICA NUCLEARE E SUBNUCLEARE

Modulo 3 : Laboratorio di analisi statistica per la fisica delle alte energie
docente: G. Sirri
http://www.unibo.it/docenti/gabriele.sirri2

Tutorial 4 : TMVA

da spedire via mail entro la prossima lezione

Per favore spedite la soluzione via mail a gabriele.sirri2@unibo.it entro la data di scadenza in modo che possiamo discutere la soluzione durante la lezione del giorno successivo.

Assicuratevi che l'oggetto della mail sia: "**Tutorial4 COGNOME1 COGNOME2**".

E' importante che tutti i vostri programmi e macro siano stati testati prima di spedirli.

Per favore commentate il codice in maniera ragionata e comprensibile.

E' raccomandato il lavoro in gruppi di due, ma **non dimenticate di dichiarare entrambi i nomi**.

Riferimenti:
- TMVA web site: http://tmva.sourceforge.net
- TMVA manual: TMVA - Toolkit for Multivariate Data Analysis , arXiv:physics/0703039v5 [physics.data-an]
- TMVA workshop (http://indico.cern.ch/conferenceDisplay.py?confId=112879)
In particolare "My tips and tricks"
http://indico.cern.ch/contributionDisplay.py?contribId=5&confId=112879
- http://www.unibo.it/docenti/gabriele.sirri2 --> Contenuti utili --> Analisi stat... Calendario e Materiale

La versione TMVA inclusa in ROOT contiene un set di utili macro :
- ROOT 5 : nella cartella `$ROOTSYS/tmva/test/`
- ROOT 6 : nella cartella `$ROOTSYS/tutorials/tmva/`

Per ogni esercizio, dovete creare un cartella di lavoro separata.
Potete **eseguire** le macro direttamente dalla cartella di lavoro, da riga di comando in questo modo:

- ROOT 5
```
root -l $ROOTSYS/tmva/test/TMVAClassification.C\(\"LD,MLP,BDT\"\)
root -l $ROOTSYS/tmva/test/TMVAGui.C
```

- ROOT 6
```
root -l $ROOTSYS/tutorials/tmva/TMVAClassification.C\(\"LD,MLP,BDT\"\)
root -l TMVA::TMVAGui()
```

Quando è richiesto di modificare qualcosa (per esempio `TMVAClassification.C`), copiate la macro nella cartella di lavoro e lavorate sulla copia. Si raccomanda di non modificare le macro presenti in `$ROOTSYS/tmva/test/` o `$ROOTSYS/tutorials/tmva/` .

Nota: in Windows la sintassi è `TMVAClassification.C\(\"LD,MLP,BDT\"\)` , senza '\' prima delle parentesi.

Scaricare i programmi (da AMS Campus tra i "Prerequisites"):

- `tmva_class_example.root` root file per l'esercizio tmva_ex70.C
  http://root.cern.ch/files/tmva_class_example.root
- `exercise71/*.*` una serie di file per l'esercizio tmva_ex71.C

Esercizio 70 – TMVAGui

Creare una cartella di lavoro "tmva_ex70"

Copiare `tmva_class_example.root` nella cartella di lavoro

Eseguire la macro TMVAClassification.C . Eseguire il training dei classificatory LD, MLP, BDT sul campione di dati di test. Non è richiesto di modificare la macro, solo di eseguirla:

```
root[] .x $ROOTSYS/tmva/test/TMVAClassification.C("LD,MLP,BDT")
```

Nel frattempo… aprire `TMVAClassification.C` e localizzare nel codice dove sono definite:

i) le variabili di input per il training;
ii) le variabili spettatori;
iii) il tree del segnale e del background;
iv) il peso del segnale e del background;
v) i tagli di selezione del segnale e del background;
vi) il numero di eventi di training e di testing;
vii) la scelta dei metodi multivariate da applicare;
viii) le chiamate di avvio del training, del test e dei paragoni fra I vara metodi.

Per usare il set di macro di TMVA (e vedere l'output del training e del test), eseguire:

```
root[] .x $ROOTSYS/tmva/test/TMVAGui.C     // se root v5
root[]  TMVA::TMVAGui()                    // se root v6
```

(Allegare i .png dei plot che ritenete più significativi)

(SKIP THE FOLLOWING)

Eseguire l'applivazione reader usando il seguente commando:
root -l TMVAClassificationApplication.C\(\"LD,BDT\"\)

Nei prossimi esercizi "XX" e "YYY" si riferiscono al vostro numero di matricola, secondo la seguente formula: NUMERO DI MATRICOLA = YYY?XX .

Esercizio 71 – Analisi multivariata FISHER

Copiare i file della cartella `exercise71` in una cartella di lavoro.

1. GENERAZIONE
Eseguire la macro **generateData.cxx** per generare ntuple di dati, i cui valori seguono una distribuzione tridimensionale (= 3 osservabili) per il segnale e un'altra per il fondo.

*Utilizzare la macro **plot.cxx** per guardare le distribuzioni.*

2. TRAINING
Eseguire la macro **tmvaTrain.cxx** per determinare i coefficienti del discriminante di Fisher. Questi coefficienti sono scritti in un file testo nella cartella **weights**. Verificare il contenuto del file e il log del comando per individuare i coefficienti.

3. ANALISI
Eseguire **analyzeData.cxx** per analizzare i dati generati.

Supponendo che le probabilità a priori di segnale e fondo siano uguali, rispondere alle seguenti domande:

a) Quali sono le efficienze per segnale e fondo se richiedete $t_{Fischer} > 0$?
b) Supponiamo siano attesi XX eventi di segnale e YYY eventi di fondo.
Qual è la purezza del segnale con questo taglio ?

Inserire dei contatori nel codice di analyzeData.cxx per rispondere a queste domande.

Scrivere una macro per visualizzare e confrontare gli istogrammi hFishSig e hFishBkg.

Si può usare come esempio la macro plot.cxx.

Esercizio 72 – Analisi multivariata Multi Layer Perceptron (MLP)

Modificare il programma **tmvaTrain.cxx** e **analyzeData.cxx** per includere una rete neutrale con uno strato nascosto con 3 nodi.

Per creare la rete neurale dovete inserire il seguente codice:
factory->BookMethod(TMVA::Types::kMLP, "MLP", "H:!V:HiddenLayers=3");
(si veda il manuale di TMVA per maggiori dettagli.

I coefficienti della rete neurale sono salvati nella cartella **weights**.

Analizzare infine i dati usando la rete neurale.

Dovete aggiungere la chiamata reader->BookMVA
usando il nome corrispondete (rimpiazzate Fisher con MLP).

Creare e riempire altri due istogrammi per guardare la distribuzione della statistica MLP per il segnale e il fondo (analogamente agli istogrammi per il discriminante di Fischer).

c) Quali sono le efficienze su segnale e fondo se si richiede $t_{MLP} > 0.5$?
d) Supponiamo siano attesi XX eventi di segnale e YYY eventi di fondo.
Qual è la purezza del segnale con questo taglio ?

Esercizio 73 – Analisi multivariata Boost Decision Tree (BDT)

Modificare il programma **tmvaTrain.cxx** e **analyzeData.cxx** per includere un Boost Decision Tree con 200 boosting iterations.

> *factory->BookMethod(TMVA::Types::kBDT, "BDT", "NTrees=200:BoostType=AdaBoost");*
> *(si veda il manuale di TMVA per maggiori dettagli.*

I coefficienti del BDT sono salvati nella cartella **weights**.

Analizzare infine i dati .

> *Dovete aggiungere la chiamata reader->BookMVA*
> *usando il nome corrispondete (rimpiazzate Fisher con BDT).*

Creare e riempire altri due istogrammi per guardare la distribuzione della statistica *BDT* per il segnale e il fondo (analogamente all'esercizio precedente).

e) Quali sono le efficienze su segnale e fondo se si richiede $t_{BDT} > 0.5$?
f) Supponiamo siano attesi XX eventi di segnale e YYY eventi di fondo.
Qual è la purezza del segnale con questo taglio ?

Esercizio 74 – Selezione ottimale

Si utilizzino le informazioni degli esercizi precedenti.

Siano date le sezione d'urto per i processi di segnale $\sigma_s = 3$ fb e di background $\sigma_B = 3$ fb .

Il campione di dati corrisponde ad una luminosità integrata $L = 20\ \text{fb}^{-1}$.

Si vogliono selezionare gli eventi di segnale richiedendo che la statistica test t sia maggiore di un certo valore t_{cut} usando il discriminante di Fisher o il MLP o il BDT.

Si trovino e si mostrino graficamente il numero di eventi attesi di segnale e background s e b in funzione di t_{cut} , facendo uso delle formule:

$$s = \sigma_s L \epsilon_s$$
$$b = \sigma_b L \epsilon_b$$

dove ϵ_s e ϵ_b sono le efficienze di segnale e di background, cioè le probabilità di accettare eventi di segnale e background usando la selezione $t > t_{cut}$.

$$\epsilon_s = P(t > t_{cut} | s)$$
$$\epsilon_b = P(t > t_{cut} | b)$$

Supponiamo di voler scegliere il valore ottimale di t_{cut} che fornisca il miglior test per la scoperta di del processo che produce il segnale. Questo può essere fatto massimizzando la *expected discovery significance*:

$$Z = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

Mostrare graficamente l'andamento di Z in funzione di t_{cut}. Si trovi il valore di t_{cut} che massimizza Z e si determini la corrispondente *expected discovery significance*.