ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN
Istituto Nazionale
di Fisica Nucleare

Analisi Statistica dei dati nella Fisica Nucl. e Subnucl. [Laboratorio 💻]

Gabriele Sirri
Istituto Nazionale di Fisica Nucleare

2015.04.30

Sommario

- Comunicazioni
- 💻 Esercizio 4
- Introduzione a RooStats
- 💻 Esercizio 5

Comunicazioni

Calendario

- Lunedì 23 febbraio 2015 14-16 M. Sioli

MARZO

- Lunedì 2 marzo 2015 14-16 M. Sioli
  Giovedì 5 marzo 2015 11-13 T. Chiarusi
- Lunedì 9 marzo 2015 14-16 M. Sioli
  Giovedì 12 marzo 2015 11-13 M. Sioli
- Lunedì 16 marzo 2015 14-16 M. Sioli
  Giovedì 19 marzo 2014 11-13 T. Chiarusi
- Lunedì 23 marzo 2015 14-16 M. Sioli
  Giovedì 26 marzo 2015 11-13 M. Sioli
  Giovedì 26 marzo 2015 16-18 G. Sirri
- Lunedì 30 marzo 2015 14-16 M. Sioli

APRILE

- Mercol. 8 aprile 2015 10-13 M. Sioli/T.Chiar.
- **Giovedì 9 aprile 2015 11-13 G. Sirri**
- Lunedì 13 aprile 2015 14-16 M. Sioli
  Giovedì 16 aprile 2015 11-13 T. Chiarusi
  Giovedì 16 aprile 2015 16-18 G. Sirri
- Lunedì 20 aprile 2015 14-16 M. Sioli
  Giovedì 23 aprile 2015 11-12 G. Sirri
  Giovedì 23 aprile 2015 12-13 T. Chiarusi

Giovedì 30 aprile 2015 11-13 G. Sirri
Giovedì 30 aprile 2015 16-18 G. Sirri

MAGGIO

- Lunedì 4 maggio 2015 14-16 M. Sioli
  Giovedì 7 maggio 2015 11-13 T. Chiarusi
- Lunedì 11 maggio 2015 14-16 M. Sioli
  Giovedì 14 maggio 2015 11-13 G. Sirri
  Giovedì 14 maggio 2015 16-18 G. Sirri
- Lunedì 18 maggio 2015 14-16 M. Sioli
  Giovedì 21 maggio 2015 11-13 T. Chiarusi
- Lunedì 25 maggio 2015 14-16 M. Sioli

Tutte le lezioni in Aula C, via Irnerio

Esercitazioni (soluzioni)

	Soluzione su AMSCampus:	Si'	Si'	NO	NO
	Lista Mail	**ES. 1**	**ES. 2**	**ES. 3**	**ES. 4**
1	sì	sì	sì	sì	
2	sì	sì	sì	sì	sì
3	sì	sì	sì	sì	
4	sì	sì	sì		
5	sì	sì	sì		
6	sì	sì	sì	sì	
7	sì	sì	sì	sì	
8	sì	sì	sì	sì	
9	sì	sì	sì	sì	
10	sì	sì	sì		
11		sì	sì		
12	sì	sì	sì	sì	
13	sì	sì	sì		
14	sì	sì	sì	sì	
15	sì	sì	sì	sì	
16	canc				

Le soluzioni di **Esercizio 3** **NON** sono ancora pubblicate in AMSCampus.

accesso riservato agli iscritti a
gabriele.sirri2.ASD-2015
con password (richiedetela via mail)

Esercizio 4

RECAP - Esercizio 4 - tmva_ex0

[0] tmva_ex0.C

- Create a working folder “tmva_ex0”
- download http://root.cern.ch/files/tmva_class_example.root to the working folder

- Run your first job using the macro TMVAClassification.C .
  Train the classifiers LD, MLP, BDT on the test data.
  You are not requested to modify the macro. You have just to run it:
  root –l $ROOTSYS/tmva/test/TMVAClassification.C\(\"LD,MLP,BDT\"\)

- Open **TMVAClassification.C** and have a look to the code. Locate where are defined :
  i) input variables for the training;
  ii) spectator variables;
  iii) signal and background trees;
  iv) signal and background weights;
  v) selection cuts on signal and background;
  vi) number of training and testing events;
  vii) booking of MVA methods;
  viii) calls to start of training, testing and method comparison

To use the TMVA collections of macros (and see the output of your training and testing), type :

root -l $ROOTSYS/tmva/test/TMVAGui.C

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

RECAP - Esercizio 4 - tmva_ex1 parte 1

[1] tmva_ex1.C (parte 1)

Lo scopo di questo esercizio è fare una semplice analisi multivariata con il pacchetto TMVA di ROOT.

Scaricate il tar file con il codice per l'esercizio da:
http://hep.fi.infn.it/ciulli/Site/Analisi Dati files/tmvaExamples.tar
quindi in una directory date il comando tar -xvf tmvaExamples.tar.

Per prima cosa usate la macro **generateData.C** per generare due n-tuple di dati, i cui valori seguono una distribuzione tridimensionale per il segnale e un'altra per il fondo. La macro **plot.C** può essere usata per guardare le distribuzioni (eseguite **root** e poi dal prompt date il comando **.x plot.C**).
Usate poi la macro **tmvaTrain.C** per determinare i coefficienti del discriminante di Fisher. Questi coefficienti sono scritti in un file nella sotto-directory weights come file testo. Guardate il log del comando e il contenuto del file per **individuare i coefficienti**.

Infine usate **analyzeData.C** per analizzare i dati generati. Supponete che le probabilità a priori di segnale e fondo siano uguali. **Quali sono le efficienze per segnale e fondo se richiedete tFisher > 0?** E qual'è la **purezza** del segnale selezionato con questo taglio?
(Modificate il codice analizeData.C inserendo dei contatori per rispondere a queste domande).

Scrivete una macro per **visualizzare e confrontare gli istogrammi hFishSig e hFishBkg**.
Potete partire come esempio dalla macro plotUniform.C del problema 1.

RECAP - Esercizio 4 - tmva_ex1 parte 2

[1] tmva_ex1.C (parte 2)

Adesso modificate il programma **tmvaTrain.cc** e **analyzeData.C** per includere una rete neurale con uno strato nascosto con 3 nodi.

Per creare la rete neurale dovete inserire la linea:

factory->BookMethod(TMVA::Types::kMLP,"MLP","H:!V:HiddenLayers=3");

dove “MLP” sta per “Multi Layer Perceptron” (si veda il manuale di TMVA per maggiori dettagli). Anche i coefficienti della rete neurale sono salvati in un file nella sottodirectory weights.

Analizzate infine i dati usando la rete neurale. Dovrete aggiungere la chiamata
reader->BookMVA;
usando il nome corrispondente (rimpiazzate Fisher con MLP).

Create e riempite altri due istogrammi per guardare la distribuzione della statistica MLP per il segnale e il fondo (analogamente agli istogrammi per il discriminante di Fisher).

Quali sono le efficienze su segnale e fondo se si richiede tMLP > 0.5?
Qual`e la purezza del segnale?

Introduzione a RooSTATS

RooStats

RooStatsTutorial_120323.pdf
https://indico.desy.de/getFile.py/access?contribId=15&resId=3&materialId=slides&confId=5065
slides da 1 a 14

Hypothesis Test with Profile Likelihood

- Profile Likelihood can be used for hypothesis tests using the asymptotic properties of the profiled likelihood ratio:

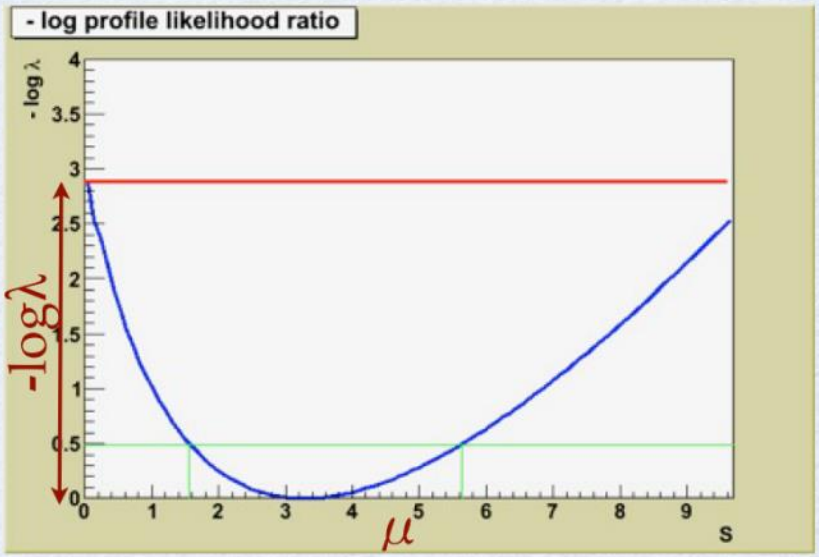
$$\lambda(\mu) = \frac{L(x|\mu, \hat{\hat{\nu}})}{L(x|\hat{\mu}, \hat{\nu})}$$

- Null hypothesis (H_0): $\mu = \mu_0$
- Alternate hypothesis (H_1): $\mu \neq \mu_0$

- Distribution of $-2\log\lambda$ is asymptotically a χ^2 distribution under H_0
- p-value and significance can then be obtained from the $-2\log\lambda$ ratio

significance: $n_\sigma = \sqrt{-2\log\lambda}$

```
// set value of POI  to zero
// one can also use model.SetSnapshot(*mu)
S->setVal(0);
plc.SetNullParameters(*mu);
HypoTestResult* hypotest =plc.GetHypoTest();
double alpha = hypotest->NullPValue();
double significance = hypotest->Significance()
```



Exercise time (Esercizio 5) 💻

RooStatsTutorial_120323.pdf
https://indico.desy.de/getFile.py/access?contribId=15&resId=3&materialId=slides&confId=5065

Esercizio 5 - roostats_ex1

[1] roostats_ex1.C
Riprendiamo l'esercizio della lezione precedente
(gaussian signal + exponential background, extended p.d.f.) .

Trovate un template già fatto qui : http://campus.unibo.it/186979/

In ***makemodel*** *:* c'è la costruzione del modello , la generazione di un set di 1000 dati simulato. Il workspace è salvato in «model.root».

Alcune modifiche rispetto alla lezione precedente :

Range e Valori iniziali dei parametri sono stati modificati.
«mean» e «sigma» sono fissati come costanti.

I dati sono importati nel workspace e salvati su file.

In ***usemodel*** : legge il modello dal file, esegue il fit del modello e disegna dati e risultato del fit in un plot.

Esercizio 5 - roostats_ex1

Si modifichi il modello:

- Specificare i componenti del modello per i tool statistici di roostat: osservabile e parametro di interesse.
- Utilizzare il numero di eventi di segnale come unico parametro di interesse.
- Fissare costanti tutti gli altri parametri del modello.
- Importare la configurazione nel workspace e salvare su file.

Si modifichi l'uso del modello:

- leggere il modelConfig dal workspace
  esempio: ModelConfig* mc = (ModelConfig*) w.obj("ModelConfig");
- calcolare un Confidence Interval utilizzando il ProfileLikelihoodCalculator
- Disegnare il profilo della likelihood e sovrapporre l'intervallo
- calcolare la discovery significance utilizzando il profilelikelihoodcalculator come test di ipotesi
- scrivere sulla console i limiti dell'intervallo e la significatività

ROOSTATS : https://twiki.cern.ch/twiki/bin/view/RooStats
short tutorial: https://twiki.cern.ch/twiki/bin/view/RooStats/RooStatsTutorialsAugust2012

Guardate la Class Reference→ *https://root.cern.ch/root/html/RooStats__ProfileLikelihoodCalculator.html*

Function Members (Methods)

public:

```
                                virtual ~ProfileLikelihoodCalculator ()
                                 static TClass* Class ()
                virtual RooStats::HypoTestResult* GetHypoTest () const
                virtual RooStats::LikelihoodInterval* GetInterval () const
                                virtual TClass* IsA () const
RooStats::ProfileLikelihoodCalculator& operator= (const RooStats::ProfileLikelihoodCalculator&)
RooStats::ProfileLikelihoodCalculator ProfileLikelihoodCalculator ()
RooStats::ProfileLikelihoodCalculator ProfileLikelihoodCalculator (const RooStats::ProfileLikelihoodCalculator&)
RooStats::ProfileLikelihoodCalculator ProfileLikelihoodCalculator (RooAbsData& data, RooStats::ModelConfig& model, Double_t
                                        size = 0.05)
RooStats::ProfileLikelihoodCalculator ProfileLikelihoodCalculator (RooAbsData& data, RooAbsPdf& pdf, const RooArgSet&
                                        paramsOfInterest, Double_t size = 0.05, const RooArgSet* nullParams = 0)
                          virtual void ShowMembers (TMemberInspector&)
                          virtual void Streamer (TBuffer&)
                                  void StreamerNVirtual (TBuffer& ClassDef_StreamerNVirtual_b)
```

Suggerimento: usate questo costruttore

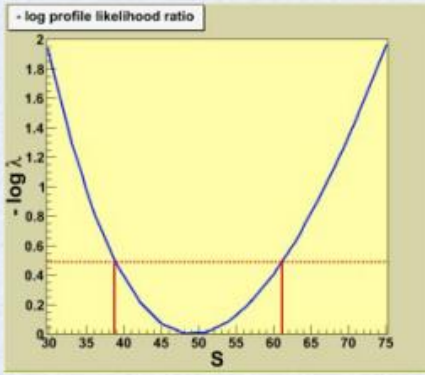
Anziché questo

```
// create class using data and model
ProfileLikelihoodCalculator plc(*data, *model, *POI);

// set the confidence level
plc.SetConfidenceLevel(0.683);

// compute the interval
LikelihoodInterval* interval = plc.GetInterval();
double lowerLimit = interval->LowerLimit(*S);
double upperLimit = interval->UpperLimit(*S);

// plot the interval
LikelihoodIntervalPlot plot(interval);
plot.Draw();
```



- For one-dimensional intervals:
  - 68% CL (1 σ) interval : $\Delta\log\lambda = 0.5$
  - 95% CL interval : $\Delta\log\lambda = 1.96$
- `LikelihoodIntervalPlot` can plot the 2D contours

Statistics School 2012, Desy 14

Ovvero :
passategli come argomento il ModelConfig
e
non singolarmente il Modello e il POI

Esercizio 5 - roostats_ex2 , roostats_ex3

[2] roostats_ex2.C
Aggiungere l'intervallo calcolato con Feldman-Cousin

suggerimento : modificare solo usemodel()

guardare il codice in : $ROOTSYS/tuturials/roostats/**IntervalExamples.C**

[3] roostats_ex3.C
Definire tau e Nb come nuisance parameters
ripetere i test dell'ex. 1 e confrontarli

suggerimento modificare makemodel() (vedere il tutorial)

ROOSTATS : https://twiki.cern.ch/twiki/bin/view/RooStats
short tutorial: https://twiki.cern.ch/twiki/bin/view/RooStats/RooStatsTutorialsAugust2012