plotgini - Dokumentation

Version 11. Oktober 2013

Kurze Dokumentation zum Programm "plotgini" Teil 1 - für Klassifizierungen

bug-reports bitte an `gunnar.klaemke@phi-tps.de`

Aufrufkonvention

```
plotgini [options] <rootInFile> <outputfile> <truth_name>
         <est_pred_names, separated by commas> [<weight_name>]
```

- **options**:
  - `-C/-c` muss angegeben werden
  - `-l/-L` für die lange Fassung inkl. S/B-Scans.
  - `-s/-S` für die Kurzform (default).
  - `-F/-f` falls die Variablen im Root-Tree als 'float' gespeichert sind (die Default-Annahme ist 'double')
- **rootInFile**: Eingabe-Datei im `.root` Format
- **outputFile**: Name der Ausgabe-Datei. Es wird `outputFile.ps` erzeugt.
- **truth_name**: Name der Variable mit der truth- bzw. target-Information.
- **est_pred_names**: eine oder mehrere Variablennamen (als Komma-separierte Liste), die jeweils einen Network-Output bezeichnen.
- **weight_name** (optional): Name der Variable, die das Ereignis-Gewicht beschreibt.

Die folgenden Ausführungen beziehen sich auf die lange Fassung.

Seiten 1 bis N

N sei die Zahl der angegebenen NB-Output-Variablen.

Oberer Plot

Dargestellt ist jeweils ein Histogramm der NB-Output-Variablen für den Signal-Fall (d.h. truth=1) in rot und einmal für den Untergrund-Fall (d.h. truth=0) in schwarz. Falls ein Gewicht angegeben wurde, werden die Ereignisse gewichtet histogrammiert.

Unterer Plot

Aufgetragen ist der NB-Output gegen die purity, d.h. dem $S/(S+B)$ Verhältnis (S=Signal, B=Untergrund) in jedem Bin des NB-Output-Histogramms. Die Darstellung ist leicht modifiziert, gemäß der Formel:

$$p'_n = p_n + \frac{N_{n+1} - N_{n-1}}{\max(N_n, N_{n+1} + N_{n-1})} \cdot \frac{1}{2 \cdot \text{Anzahl Bins}}$$

Diese soll die Verteilung der Ereignisse innerhalb des Bins abschätzen und in erster Näherung auf die Bin-Mitte korrigieren.
Dabei ist:

- p_n der purity-Wert im bin N
- p'_n der modifizierte purity-Wert im Bin N, wie er im Plot eingezeichnet wird
- N_n der Eintrag im bin n des Signal-Histogramms
- Anzahl Bins ist im allg. gleich 100

In diesem Plot sollten möglichst alle Punkte auf der Diagonalen liegen, nur dann kann der Netzwerk-Output direkt als Wahrscheinlichkeit interpretiert werden.

Seite N+1

Oberer Plot

Für jede angegebene NB-Output-Variable ist die Signal-efficiency in Abhängigkeit von der Signal-purity für zwei verschiedene Schnitt-Verläufe dargestellt.

- Signal-efficiency ist definiert als: $\frac{\text{Anzahl Signal Ereignisse nach Schnitt}}{\text{Anzahl Signal Ereignisse insgesamt}}$
- Signal-purity ist definiert als: $\frac{\text{Anzahl Signal Ereignisse nach Schnitt}}{\text{Anzahl Signal+Untergrund Ereignisse nach Schnitt}}$

Die zwei Schnitt-Verläufe sind (bezogen auf die oberen Plots der Seiten 1 bis N):

- von links nach rechts, NB-Network-Output $>$ Schnitt
- von rechts nach links, NB-Network-Output $<$ Schnitt

Unterer Plot

Das gleiche wie im oberen Plot, nur dass die Rollen von Signal- und Untergrund vertauscht sind. Also:

- Background-efficiency ist definiert als: $\frac{\text{Anzahl Untergrund Ereignisse nach Schnitt}}{\text{Anzahl Untergrund Ereignisse insgesamt}}$
- Background-purity ist definiert als: $\frac{\text{Anzahl Untergrund Ereignisse nach Schnitt}}{\text{Anzahl Signal+Untergrund Ereignisse nach Schnitt}}$

Seite N+2 (Gini Index)

Oberer Plot (Gini-Index für Signal)

Für jede angegebene NB-Output-Variable:
Dargestellt ist die Signal-efficiency in Abhängigkeit von der efficiency (die Daten ergeben sich aus den Schnitt-Verläufen von Seite N+1). Dies ist der sogenannte Lift-Chart.

- efficiency ist definiert als: $\frac{\text{Anzahl Signal+Untergrund Ereignisse nach Schnitt}}{\text{Anzahl Signal+Untergrund Ereignisse insgesamt}}$
- Die Diagonale entspricht dem Grenzfall, wenn der NB-Output eine exakt gleiche Verteilung für Signal und Untergrund liefert, also keine Trennung möglich ist.
- Die obere linke Gerade entspricht dem Grenzfall, wenn der NB-Output eine perfekte Trennung von Signal und Untergrund ermöglicht.
- Der maximal mögliche Gini-Index ist die Fläche des aufgespannten Dreiecks.
- Der Gini Index ist die Fläche, die von der Kurve und der Diagonalen eingeschlossen wird.

Unterer Plot (Gini-Index für Untergrund)

Das gleiche wie im oberen Plot, nur dass die Rollen von Signal- und Untergrund vertauscht sind.

Seite N+3 (Gini Index im 1:1 Fall)

Im Allgemeinen ist die Anzahl von Signal- und Untergrund-Ereignissen im vorliegenden Sample unterschiedlich. Dargestellt ist der Gini-Index für Signal und Untergrund, wie auf Seite N+2, aber für den Fall, dass Anzahl Signalereignisse=Anzahl Untergrundereignisse ist, d.h. $S/B = 1:1$. Für die Berechnung wurden die NB-Output-Verteilungen für Signal- und Untergrund (Seiten 1 bis N) vom ursprünglichen S/B-Verhältnis auf ein Verhältnis $S'/B' = 1$ umgewichtet. (Eine einfache Skalierung der Histogramme.) Die Berechung der benötigten Gewichte verläuft wie folgt:

$$\begin{aligned} r &\equiv \frac{S'/B'}{S/B} \\ w_S &= \frac{N_S + N_B}{N_S + N_B \cdot \frac{1}{r}} \\ w_B &= \frac{N_S + N_B}{N_S \cdot r + N_B} \end{aligned} \tag{1}$$

mit:

- S/B: das Signal- zu Untergrundverhältnis im vorliegenden Sample
- S'/B': das Ziel-S/B-Verhältnis, in diesem Falle $=1:1$
- w_S: das Gewicht, das ein Signal-Ereignis bekommt

- w_B: das Gewicht, das ein Untergrund-Ereignis bekommt
- N_S: die Summe der Gewichte aller Signal-Ereignisse im vorliegenden Sample (entspricht der Anzahl der Signal-Ereignisse, falls keine Gewichtung vorliegt)
- N_B: die Summe der Gewichte aller Untergrund-Ereignisse im vorliegenden Sample (entspricht der Anzahl der Untergrund-Ereignisse, falls keine Gewichtung vorliegt)

Diese Art der Umgewichtung garantiert, dass die Gesamtsumme der Gewichte unverändert bleibt:

$$N_S \cdot w_S + N_B \cdot w_B = N_S + N_B$$

Seite N+4 (Signal-efficiency vs. Untergrund-efficiency)

Für jede angegebene NB-Output-Variable ist die Signal-efficiency gegen die Untergrund-efficiency aufgetragen. Die Daten kommen aus den bereits berechneten Schnitt-Verläufen. Der obere Plot hat eine lineare, der untere eine logarithmische Skala für die Untergrund-efficiency.

Seite N+5 (Likelihood und R^2-Wert)

Oberer Plot

Berechnet wird die Likelihood $\mathcal{L}$ und die inklusive Likelihood $\mathcal{L}_{incl}$ für verschiedenen S'/B'-Werte gemäß der folgenden Formeln. Die Werte, die dem S/B-Verhältnis im Sample entsprechen, sind als Punkte in der Grafik markiert.

$$\begin{aligned}
\log \mathcal{L} &= \frac{1}{N_S + N_B} \left(\sum_{\text{Signal},i} W_i' \cdot \log p_i' + \sum_{\text{Untergrund},i} W_i' \cdot \log(1 - p_i') \right) \\
\log \mathcal{L}_{incl} &= p_{incl}' \log p_{incl}' + (1 - p_{incl}') \log(1 - p_{incl}') \\
p_{incl}' &= \frac{S'/B'}{1 + S'/B'} \\
p_i' &= \frac{1}{1 + \left(\frac{1}{p_i} - 1\right) \cdot \frac{S/B}{S'/B'}} \quad \Leftarrow \quad \frac{p_i'}{1 - p_i'} = \frac{p_i}{1 - p_i} \cdot \frac{S'/B'}{S/B} \quad \Leftrightarrow \quad \left(\frac{s}{b}\right)'_i = \left(\frac{s}{b}\right)_i \cdot \frac{S'/B'}{S/B} \\
W_i' &= \begin{cases} W_i \cdot w_S \,, \text{ für ein Signal-Ereignis} \\ W_i \cdot w_B \,, \text{ für ein Untergrund-Ereignis} \end{cases} \qquad (2)
\end{aligned}$$

Mit

- N_S, N_B, w_S, w_B, S/B, S'/B' wie in Gl. (1).
- W_i ist das Original-Ereignisgewicht (falls im plotgini-Aufruf angegeben, sonst =1).
- p_i ist der NB-Output für das Ereignis.
- p_i' ist der auf das neue S'/B'-Verhältnis umgerechnete NB-Output für das Ereignis.
- p_{incl}' ist die inklusive purity, die dem S'/B'-Verhältnis entspricht.

Unterer Plot

Berechnet wird der R^2-Wert für verschiedene S'/B'-Werte gemäß der folgenden Formeln. Der R^2-Wert gibt an, welcher Bruchteil der Varianz durch das NeuroBayes®-Modell erklärt werden kann.

$$\begin{aligned} R^2 &= 1 - \frac{\chi^2_{indiv}}{\chi^2_{incl}} \\ \chi^2_{indiv} &= \frac{1}{N_S + N_B} \sum_{\text{Ereignisse},i} W'_i \cdot (p'_i - \tau_i)^2 \\ \chi^2_{incl} &= \frac{1}{N_S + N_B} \sum_{\text{Ereignisse},i} W'_i \cdot (p'_{incl} - \tau_i)^2 \end{aligned} \tag{3}$$

Dabei ist τ_i die truth-Information des Ereignisses i.

Seite N+6 (εD^2-Wert und Gini-Index)

Oberer Plot

Berechnet wird der εD^2-Wert für verschiedene S'/B'-Verhältnisse gemäß folgender Formel:

$$\varepsilon D^2 = \frac{1}{N_S + N_B} \sum_{\text{Ereignisse},i} W'_i \cdot (1 - 2p'_i)^2 \tag{4}$$

εD^2 ist ein Maß für die Effizienz $\times$ Dilution^2 (Verwässerung), welches in Likelihood-Fits von Teilchen/Antiteilchen-Klassifizierungen eingeht.

Unterer Plot

Dargestellt ist der Verlauf des Gini-Indexes für verschiedene S'/B'-Verhältnisse. Die Berechnung erfolgt über umgewichtete NB-Output-Histogramme für Signal und Untergrund entsprechend der Diskussion zu Gl. (1). Die gestrichelte schwarze Kurve beschreibt den Verlauf des maximal möglichen Gini-Indexes.

Teil 2 - für Wahrscheinlichkeitsdichten

Aufrufkonvention

```
plotgini [options] <rootInFile> <outputfile> <truth_names, separated by commas>
                   <mean_names, separated by commas>
                   <median_names, separated by commas> [<weight_name>]
```

- **options**:
  - `-D/-d` muss angegeben werden
  - `-F/-f` falls die Variablen im Root-Tree als 'float' gespeichert sind (die Default-Annahme ist 'double')
- **rootInFile**: Eingabe-Datei im `.root` Format
- **outputFile**: Name der Ausgabe-Datei. Es wird `outputFile.ps` erzeugt.
- **truth_names**: eine oder mehrere Variablennamen (als Komma-separierte Liste), die die Truth- bzw. das Target beschreiben.
- **mean_names**: Liste von Variablennamen, die den Mean der Netzwerk-Vorhersage angeben. Die Reihenfolge muss der Liste der angegebenen truth-Variablen entsprechen.
- **median_names**: Liste von Variablennamen, die den Median der Netzwerk-Vorhersage beschreiben. Die Reihenfolge muss ebenfalls der Liste der angegebenen truth-Variablen entsprechen.
- **weight_name** (optional): Name der Variable, die das Ereignis-Gewicht beschreibt.

Für jedes der angegebenen Tupel (truth,mean,median) wird nun folgender Output erzeugt.

Wahrscheinlichkeitsdichte

Es wird die Verteilung der Wahrscheinlichkeitsdichte (probability density distribution) für jeweils die Truth, den Mean und Median geplottet. Oben mit linearer, unten mit logarithmischer Skala.

Diagonal-Plot

Dargestellt sind zwei Profil-Histogramme über den Truth vs. Mean und den Truth vs. Median Zusammenhang.

- Die Dreiecks-Marker geben den Mittelwert (mean) der Truth im jeweiligen Bin an.
- Die Stern-Marker geben den Median der Truth im jeweiligen Bin an.
- Die Fehlerbalken geben das 68%-Konfidenzintervall um den Median herum an.

Des Weiteren wird der R^2- und der R^{MAD}-Wert berechnet, gemäß folgender Formeln:

$$\chi^2_{indiv} = \sum_i (t_i - \hat{m}_i)^2 \tag{5}$$

$$\chi^2_{incl} = \sum_i (t_i - \hat{t})^2 \tag{6}$$

$$\mathrm{MAD}_{indiv} = \sum_i |t_i - \bar{m}_i| \tag{7}$$

$$\mathrm{MAD}_{incl} = \sum_i |t_i - \bar{t}| \tag{8}$$

$$R^2 = 1 - \frac{\chi^2_{indiv}}{\chi^2_{incl}} \tag{9}$$

$$R^{MAD} = 1 - \frac{\mathrm{MAD}_{indiv}}{\mathrm{MAD}_{incl}} \tag{10}$$

Dabei ist t_i die Truth-Information, $\hat{m}_i$ der Mean, $\bar{m}_i$ der Median des Ereignisses i. Weiterhin ist $\hat{t}$ der Mean und $\bar{t}$ der Median der inklusiven Truth-Verteilung.

Lorenz-Kurve und Gini-Index

Für die Truth-, Mean- und Median-Verteilung wird die Lorenz-Kurve und der Gini-Index berechnet. Für die Berechnung der Lorenz-Kurven werden die Truth-, Mean- und Median-Werte zunächst aufsteigend sortiert und anschließend kumuliert. Jeder Punkt ($x\%$, $y\%$) entlang der Lorenz-Kurve liefert eine Aussage der Form:
Die oberen x% aller Ereignisse machen einen Anteil von y% an der Gesamtsumme aller Werte in der Verteilung aus.
Der Gini-Index $\epsilon\,[0, 1]$ ist der doppelte Inhalt der Fläche, die von der Lorenz-Kurve und der Diagonalen eingeschlossen wird.

- Der Gini-Index ist minimal ($= 0$), wenn perfekte Gleichverteilung vorliegt, d.h. alle Ereignisse haben den selben Wert.
- Der Gini-Index ist maximal ($= 1$), wenn perfekte Ungleichverteilung vorliegt, d.h. alle Ereignisse bis auf eines haben den Wert 0 und nur ein Ereignis hat einen Wert ungleich null (und hat damit einen Anteil von 100% an der Gesamtsumme aller Werte).