L'implementazione delle ROC Curve nei modelli GAMLSS come strumento di previsione per i Big Data

Paolo Mariani, Andrea Marletta

Dipartimento di Economia, Metodi Quantitativi
e Strategie di Impresa,
Università degli Studi Milano-Bicocca

Outline

Big Data

Caratteristiche dei Big Data:

- non sono necessariamente più grandi e più veloci di altri dati;
- cambiano gli oggetti della conoscenza;
- offrono un nuovo modo di intendere la ricerca;
- non sono necessariamente dati migliori;
- difficoltà di interpretazione del contesto.

Il modello delle 7 V per i Big Data

I Big Data compongono un insieme gigantesco e in continuo accrescimento, istante per istante, in ogni luogo del mondo, generato spesso inconsapevolmente da milioni di persone, con riferimento a tantissime tematiche.

7 V	**Small Data**	**Big Data**
Volume	Megabyte 10^6	Zettabyte 10^{21}
Velocità	Non in tempo reale	Tempo reale
Varietà	Strutturati	Strutturati e non
Valore	Elevato	Da dimostrare
Veridicità	Elevata	Contenuta
Validità	Elevata	Limitata
Visualizzazione	Elevata	Contenuta

Modelli GAMLSS (Rigby and Stasinopoulos, 2001)

- Generalized Additive Models for Location Scale and Shape (GAMLSS) sono modelli semi-parametrici
- Una possibile soluzione ai limiti di GLM e GAM
- Possibilità di assumere una distribuzione di probabilità per la variabile risposta non appartenente alla classe esponenziale ma ad una famiglia di distribuzione più generale $P(\mu, \hat{\phi})$
- Espansione del modello a parametri di forma e scala quali asimmetria e curtosi

Modelli GAMLSS (Rigby and Stasinopoulos, 2001)

Data una variabile risposta $y^T = (y_1, y_2, \ldots, y_n)$ e date $g_p(.)$ link function, per $p = 1, 2, 3, 4$, che collegano i parametri di distribuzione alle variabili esplicative, si ha:

$$g_p(\theta_p) = \eta_p = X_p\beta_p + \sum_{j=1}^{J_p} Z_{jp}\gamma_{jp}$$

Ogni parametro di distribuzione $(\theta_p) = (\mu, \sigma, \nu, \tau)$ può essere modellato come una combinazione di funzioni lineari e non del set di esplicative e/o l'introduzione di effetti casuali.

I vettori parametrici β_p and γ_{jp} possono essere stimati massimizzando una verosimiglianza penalizzata attraverso due algoritmi: RS o CG.

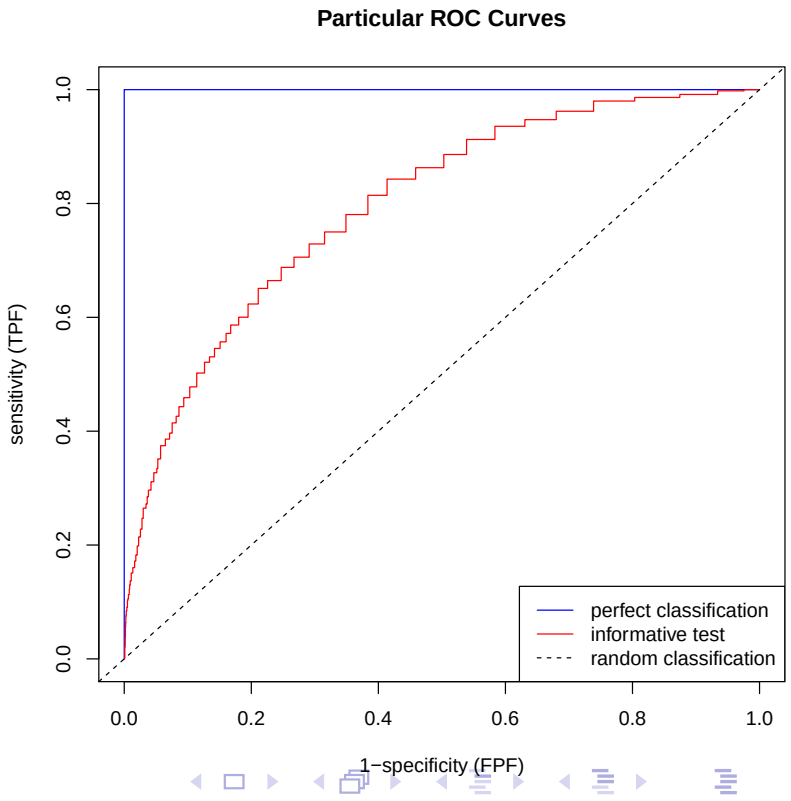
La curva ROC (Receiver Operative Characteristic)

Sia D una variabile binaria che denota la presenza/assenza di un carattere e Y il risultato del test diagnostico binario. Il grafico di FPF vs TPF è chiamato curva ROC (Sullivan Pepe, 2000). Il potere discriminatorio dei valori previsti del modello è quantificato dall'area sottesa alla curva (AUC).

	D=0	D=1
Y=0	TN	FN
Y=1	FP	TP
Totale	TN + FP	FN + TP

$$FPF = \frac{FP}{TN + FP}$$

$$TPF = \frac{TP}{TP + FN}$$



Proposta: implementazione delle ROC curve in GAMLSS

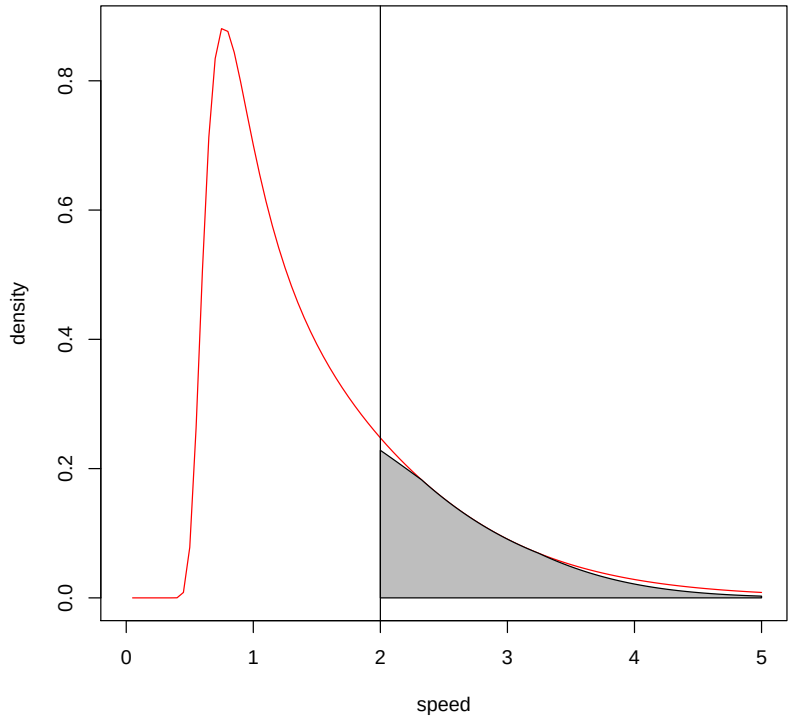
Nei modelli GAMLSS le probabilità previste $\hat{p}$ sono ottenute usando la funzione di densità della distribuzione di probabilità assunta per la variabile risposta ad una soglia prestabilita α:

$$\hat{p} = 1 - F_y(\alpha|\mu = \hat{\mu}, \sigma = \hat{\sigma}, \nu = \hat{\nu}, \tau = \hat{\tau})$$

$$\hat{p} = P(Y > \alpha) =$$

$$= 1 - P(Y \leq \alpha) =$$

$$= 1 - F_y(\alpha|\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$$



Procedura per implementare le ROC curve in GAMLSS

- Dividere il dataset in training e validation set
- Stimare un modello GAMLSS pesato con pesi $w = 1$ per il training set e $w = 0$ per il validation set
- Estrarre i valori previsti $\hat{Y}$ e trasformarli in $\hat{p}$ per ognuno dei campioni del validation set

$$\hat{Y} \rightarrow \hat{p} \qquad\qquad \hat{p} = 1 - F_y(\alpha|\mu = \hat{\mu}, \sigma = \hat{\sigma}, \nu = \hat{\nu}, \tau = \hat{\tau})$$

- Calcolare specificità e sensibilità e disegnare la curva ROC usando il valore vero D e le probabilità previste $\hat{p}$

Applicazione su dati Twitter

Una semplice applicazione considera i Tweet nell'anno 2016 dell'account ufficiale dell'autodromo nazionale di Monza ($n = 737$). I modelli GAMLSS sono proposti per misurare la relazione fra:

Variabile risposta

- Numero di like ricevuti (♡)

Set di variabili esplicative

- Numero di hashtag presenti (#)
- Numero di tag presenti (@)
- Numero di link presenti (http)

Previsione sui dati provenienti da Twitter

L'approccio proposto si pone l'obiettivo di prevedere il numero di like ricevuti in base alla presenza di alcuni caratteri speciali presenti all'interno del tweet.

In particolare, la soglia α è fissata per semplicità a 1, quindi la divisione sarà fra tweet che hanno ricevuto almeno un like e non.

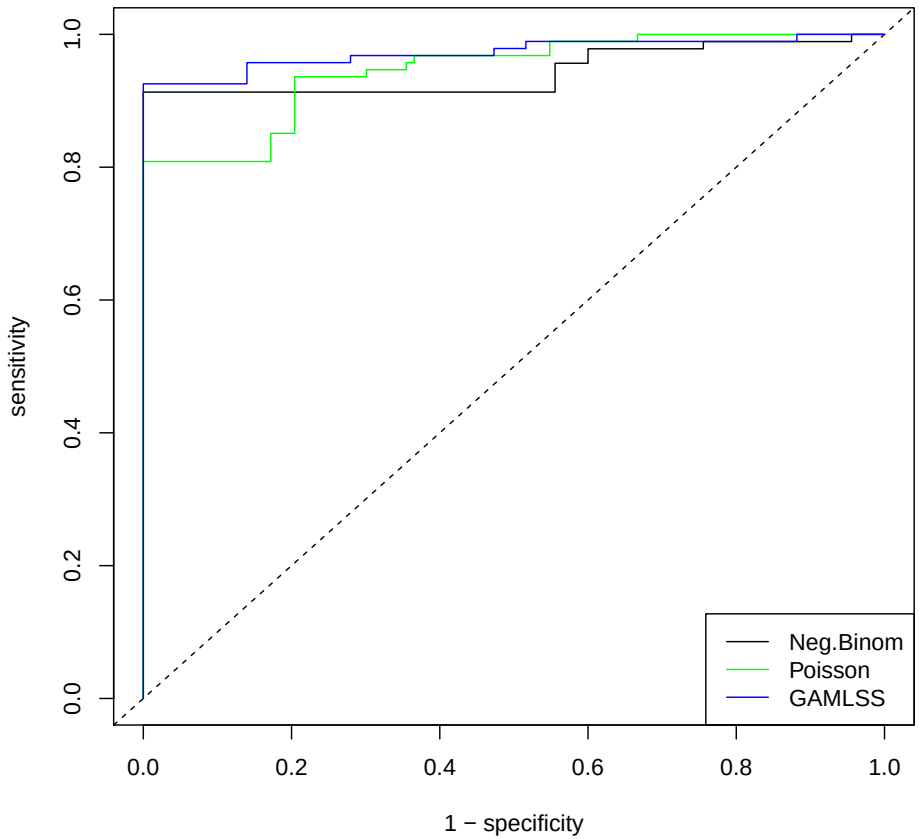
Per il modello selezionato si assume che il numero di like segua una distribuzione di Sichel (Sichel, 1973).

Numero di like	$D = 0$	$D \geq 1$	Totale
Frequenza	391	346	737

Confronto fra ROC curve per differenti modelli statistici

Poiché il numero di like è una variabile di conteggio, per il confronto con i GAMLSS sono proposti un modello di regressione di Poisson e uno con risposta Binomiale Negativa.

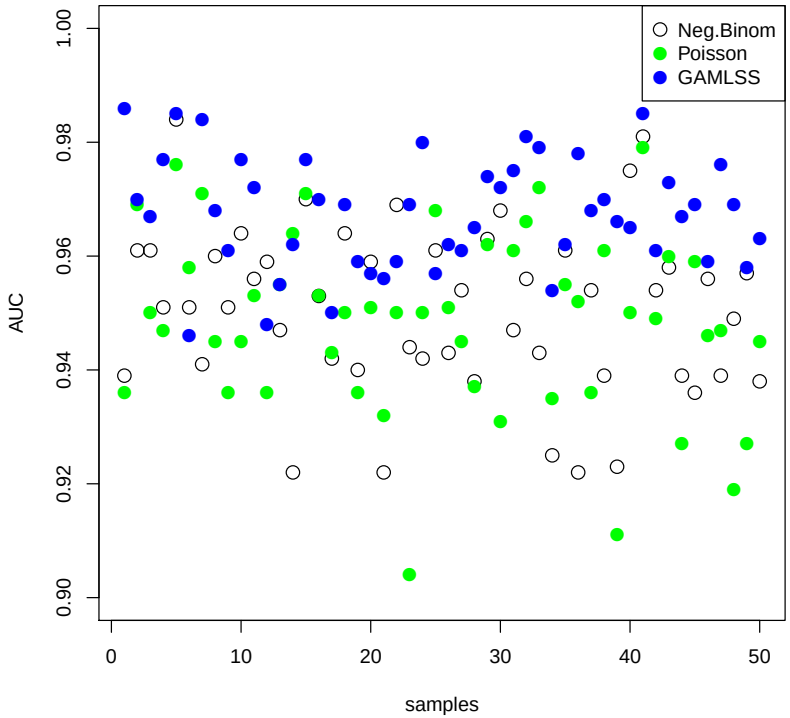
- La GAMLSS ROC curve è più alta delle altre
- Ogni curva è sopra la retta diagonale quindi il test è informativo
- Per bassi valori di specificità, le ROC curve si intersecano

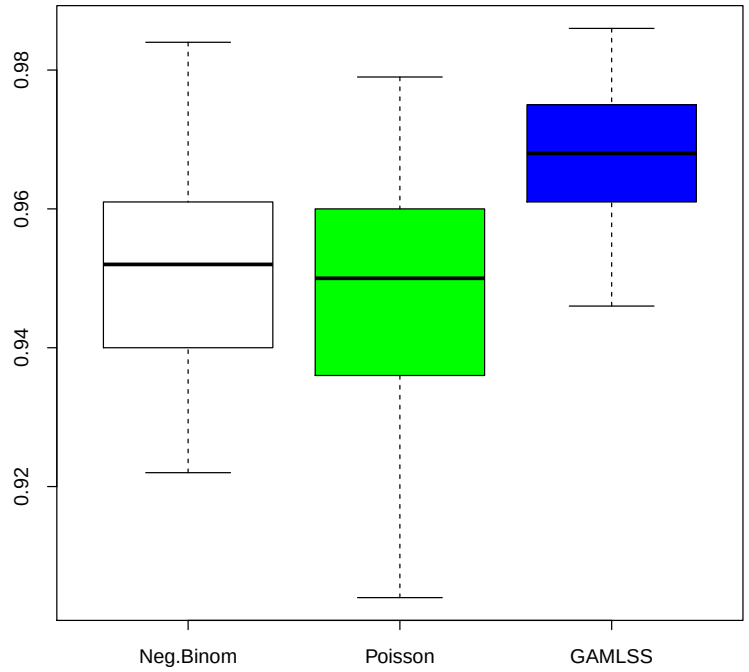


Confronto fra indici AUC per differenti modelli statistici

Dividendo il dataset in training (75%) e validation set (25%), 50 differenti campioni sono stati generati

Modello	Neg. Binom	Poisson	**GAMLSS**
$\overline{AUC}$	0.951	0.949	**0.967**





Conclusioni

- Scelta dei GAMLSS come modelli di previsione per i Big Data
- Implementazione delle ROC Curve in GAMLSS
- Possibile il confronto grafico o in termini di AUC con altri modelli statistici

Sviluppi futuri

- Previsioni con GAMLSS con aggiunta di termini non lineari
- Altre applicazioni con altre distribuzioni di probabilità