UNIVERSITY OF SALERNO

DEPARTMENT OF INFORMATION ENGINEERING, ELECTRICAL ENGINEERING AND APPLIED MATHEMATICS

Relazione Data Analysis A-H

---

prof. Postiglione Fabio
prof. Matta Vincenzo

Gruppo 06

Studente	Matricola
Alberti Andrea	0622702370
Attianese Carmine	0622702355
Capaldo Vincenzo	0622702347
Esposito Paolo	0622702292

Anno Accademico 2024 – 2025

CONTENTS

CHAPTER 1

REGRESSIONE

La regressione è un task supervisionato, questo significa che c'è bisogno di un dataset contenente le variabili indipendenti (i regressori) e la variabile dipendente. Il dataset fornito è formato da 160 campioni e da 1 regressore.

1.1 Punto 1

Prima di eseguire qualsiasi analisi è fondamentale eseguire un pre-processing dei dati. In particolare, è importante eliminare eventuali campioni che contengono dei valori mancanti. Inoltre, dal momento che il dataset fornito è unico, è necessario dividere il dataset in due ulteriori dataset:

- training-set: dati utilizzati per individuare i regressori ottimali (75%);
- test-set: dati utilizzati per testare il modello (25%).

Per stabilire il grado p del polinomio completo in X che minimizza l'errore di predizione sul test set è necessario calcolarne l'MSE di ogni polinomio. Il vincolo è quello di esplorare polinomi con grado massimo $p = 10$. La tabella 1.1 mostra i polinomi da analizzare:

Modello	**Espressione**
Modello 1	$Y = \beta_0 + \beta_1 X$
Modello 2	$Y = \beta_0 + \beta_1 X + \beta_2 X^2$
Modello 3	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
Modello 4	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$
Modello 5	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5$
Modello 6	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6$
Modello 7	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \beta_7 X^7$
Modello 8	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \beta_7 X^7 + \beta_8 X^8$
Modello 9	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \beta_7 X^7 + \beta_8 X^8 + \beta_9 X^9$
Modello 10	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \beta_7 X^7 + \beta_8 X^8 + \beta_9 X^9 + \beta_{10} X^{10}$

Table 1.1: Modelli polinomiali fino al polinomio di grado 10.

Il polinomio da scegliere è quello che minimizza la predizione sul test set. La figura 1.1 mostra che il polinomio migliore è quello di grado 4.

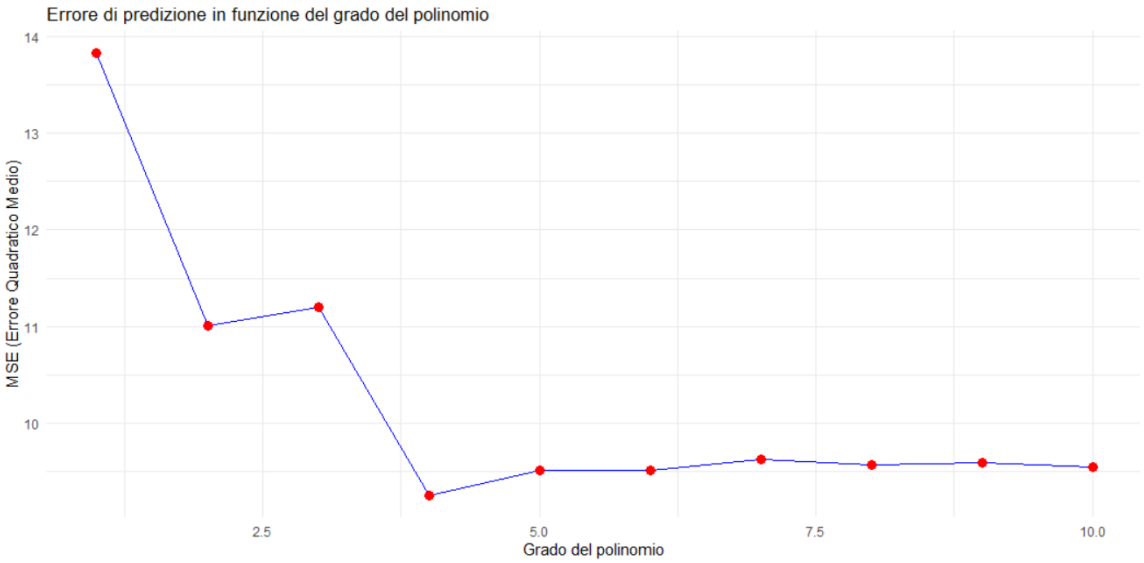

Figure 1.1: MSE su test set al variare del grado del polinomio.

1.2 Punto 2

La strategia *backward selection* prevede un processo iterativo: si parte da un modello composto da tutti i regressori, che in questo caso è un polinomio di grado 4, e ad ogni iterazione si elimina il regressore meno significativo. Il modello migliore è quello che minimizza il *Bayesian Information Criterion (BIC)*. Di seguito è riportato il summary del modello scelto.

```
Call:
lm(formula = Y ~ X2 + X4, data = d_train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1195 -1.8788  0.1707  1.8530  6.6547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1451     0.4880   2.346   0.0206 *
X2           -1.8698     0.7237  -2.584   0.0110 *
X4            1.2080     0.1956   6.177  9.8e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.8 on 117 degrees of freedom
Multiple R-squared:  0.5882, Adjusted R-squared:  0.5812
F-statistic: 83.57 on 2 and 117 DF,  p-value: < 2.2e-16
```

1.3 Punto 3

In figura 1.2 sono mostrate in nero le osservazioni, in rosso (linea continua) la funzione di regressione, in rosso (linea tratteggiata) l'intervallo di confidenza e in verde l'intervallo di predizione. Gli intervalli sono stati calcolati utilizzando un $\alpha = 0.05$. Come ci si aspettava, dal momento che l'intervallo di predizione considera anche la variabilità dei dati, è più ampio dell'intervallo di confidenza.

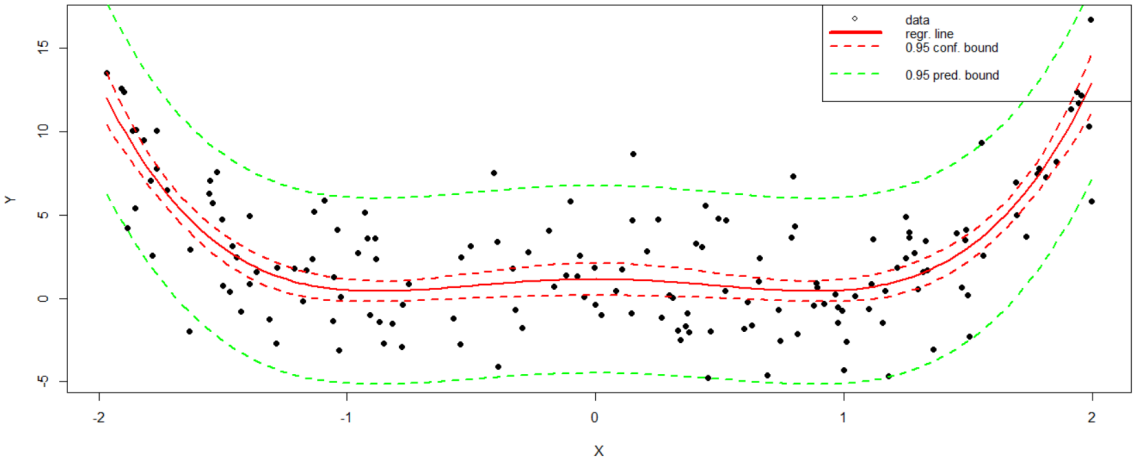

Figure 1.2: Plot dei dati, funzione di regressione e intervalli di confidenza e di predizione.

1.4 Punto 4

In figura 1.3 sono mostrati i grafici diagnostici dei residui del modello di regressione polinomiale ricavato al punto 2. Dalla figura 1.3 è possibile estrarre le seguenti informazioni:

- Residual vs Fitted: i residui sono molto vicini a 0, per cui il modello effettua una buona predizione.
- Normal Q-Q: il grafico mostra una retta, per cui l'errore si può considerare gaussiano.
- Scale-Location: non sono presenti outliers.
- Residual vs Leverage: non sono presenti punti di leva.

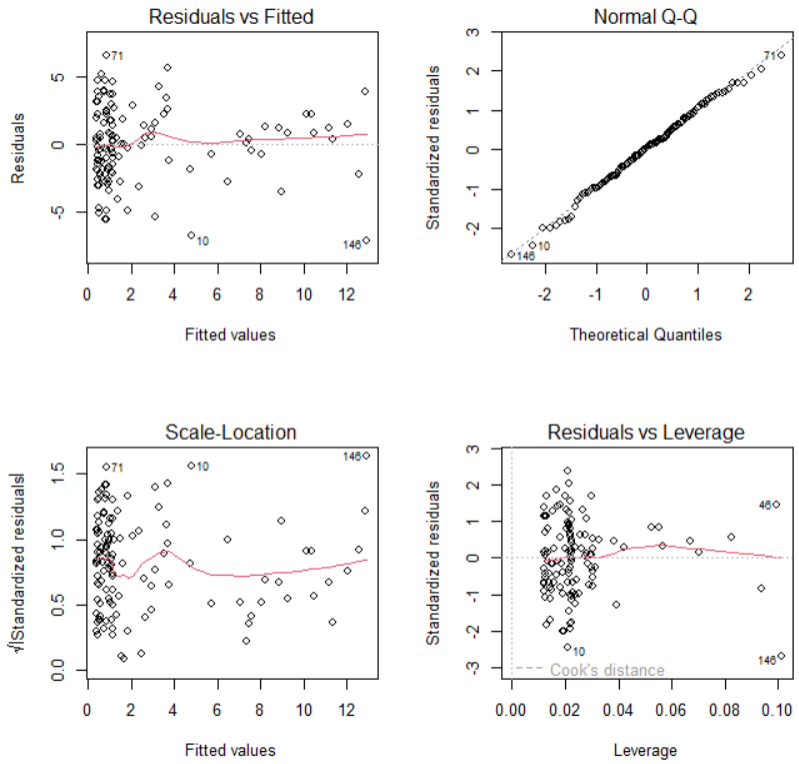

Figure 1.3: Grafici di diagnostica.

CHAPTER 2

CLASSIFICAZIONE

Il problema considerato è un caso di classificazione binaria, in cui la variabile aleatoria Y assume due possibili valori $Y \in \{a, b\}$, mentre le osservazioni $X = (x_1, x_2)$ appartengono a $\mathbb{R}^2$. Le componenti x_1 e x_2 sono indipendenti tra loro e, dato Y, hanno densità di verosimiglianza (*likelihood*) normali:

$$\ell(x_1|Y=a) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_1-0.5)^2}{2\sigma_1^2}\right)$$

$$\ell(x_2|Y=a) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_2-0.5)^2}{2\sigma_2^2}\right)$$

$$\ell(x_1|Y=b) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_1+0.5)^2}{2\sigma_1^2}\right)$$

$$\ell(x_2|Y=b) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x_2+0.5)^2}{2\sigma_2^2}\right)$$

con $\sigma_1^2 = 1$ e $\sigma_2^2 = 2$.
Le probabilità a priori delle classi sono *uniformi*, ossia:

$$P(Y=a) = P(Y=b) = \frac{1}{2}.$$

2.1 Classificatore ottimo

La regola di decisione ottima è ottenuta massimizzando la probabilità a posteriori, secondo il criterio MAP (*Maximum A Posteriori*):

$$\hat{Y} = \arg\max_{y \in \{a,b\}} P(Y=y|X=x).$$

Applicando il teorema di Bayes, la decisione ottima si riduce a confrontare le verosimiglianze ponderate dalle probabilità a priori:

$$\hat{Y} = \begin{cases} a, & \text{se } \ell(x_1|Y=a)\ell(x_2|Y=a) > \ell(x_1|Y=b)\ell(x_2|Y=b) \\ b, & \text{altrimenti.} \end{cases}$$

Di seguito, in figura 2.1, sono riportati i passaggi analitici che portano alla derivazione della regola di decisione ottima, illustrati nell'immagine seguente.

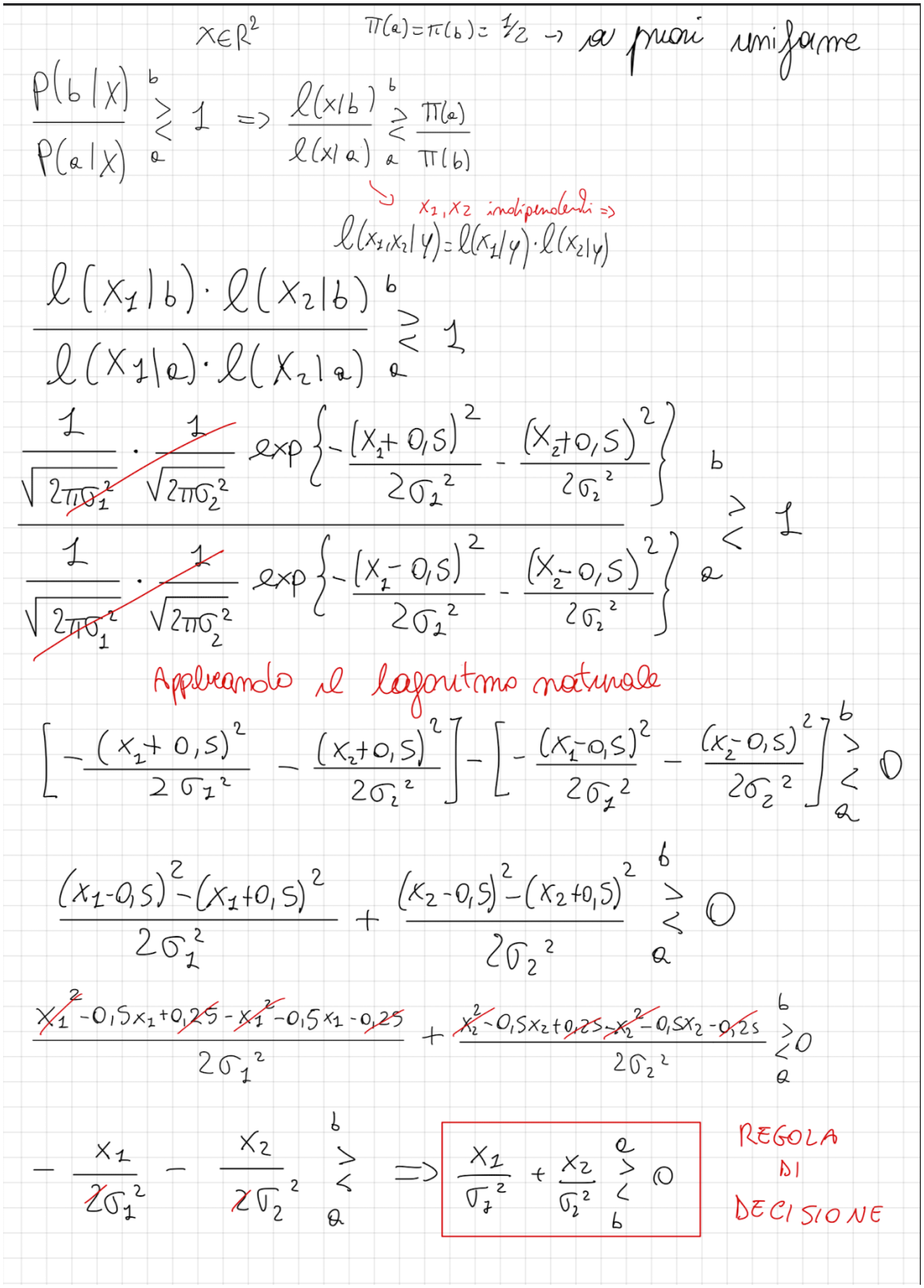

Figure 2.1: Classificatore Ottimo.

La regola di decisione ottima implementata nel codice è la seguente:

$$\hat{Y} = \begin{cases} a, & \text{se } \frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} > 0 \\ b, & \text{se } \frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} < 0 \\ \textit{scelta casuale tra le due classi}, & \text{se } \frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} = 0. \end{cases}$$

Nel caso in cui la funzione discriminante sia esattamente zero, la decisione viene presa scegliendo a o b con probabilità pari a $\frac{1}{2}$. Nel codice consegnato, la classe a è rappresentata con la label *+1*, mentre la classe b con la label *-1*.

2.2 Classificatore logistico

Il regressore logistico è un modello di classificazione che stima la probabilità di appartenenza a una classe, dato un campione $x \in \mathbb{R}^2$, con la seguente funzione di decisione:

$$P(Y = a|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

dove $\beta_0, \beta_1, \beta_2$ sono i coefficienti del modello da apprendere dai dati.
La regola di decisione del classificatore logistico è basata sul confronto con una soglia fissata a 0.5:

$$\hat{Y} = \begin{cases} a, & \text{se } P(Y = a|X) > 0.5 \\ b, & \text{se } P(Y = a|X) < 0.5 \\ \textit{scelta casuale tra le due classi}, & \textit{se } P(Y = a|X) = 0.5 \end{cases}$$

La stima dei parametri $\beta_0, \beta_1, \beta_2$ è stata effettuata mediante due approcci di addestramento differenti:

1. addestramento con *tasso di apprendimento fisso*;
2. addestramento con *tasso di apprendimento variabile*.

L'algoritmo del gradiente stocastico è stato eseguito utilizzando un dataset di addestramento composto da 8000 campioni e 10 iterazioni Monte Carlo, al fine di stimare i coefficienti $\beta_0, \beta_1, \beta_2$ tramite la media delle soluzioni ottenute nelle diverse simulazioni.

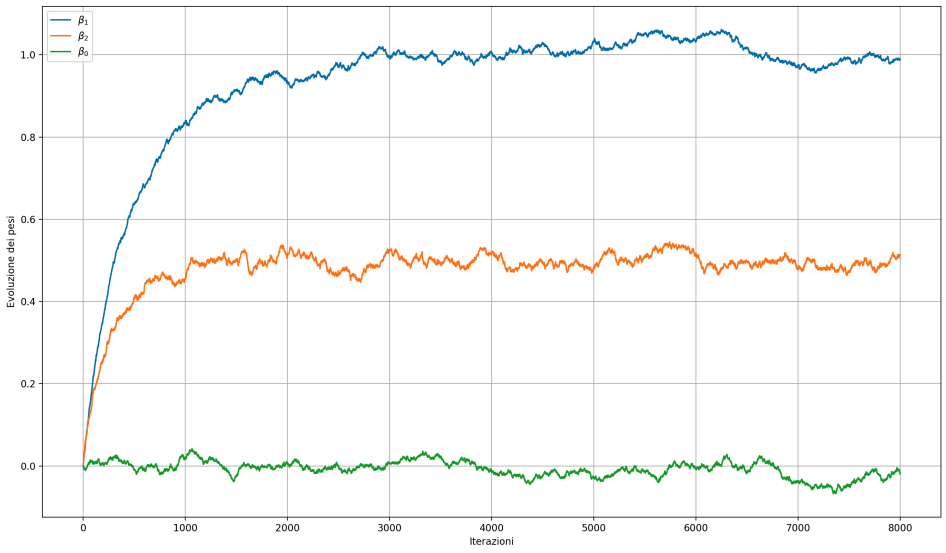

Figure 2.2: Addestramento con learning rate fisso.

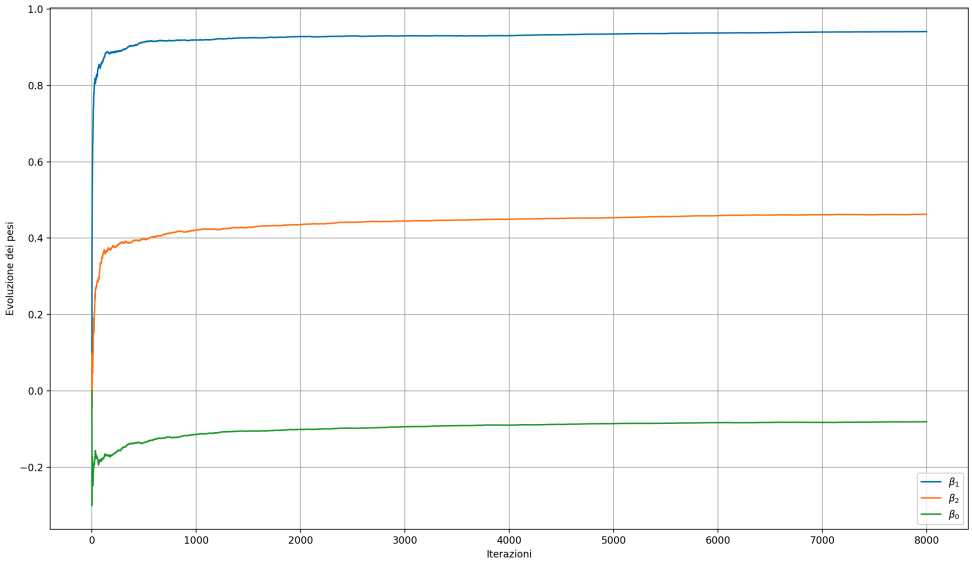

Figure 2.3: Addestramento con learning rate variabile.

2.3 Prestazioni dei modelli

I test sono stati condotti utilizzando un set di 2000 campioni e 10 iterazioni Monte Carlo, generando un test set diverso per ogni iterazione. I risultati ottenuti mostrano che le probabilità di decisione corretta per il classificatore ottimo e per il classificatore basato sulla regressione logistica (sia con learning rate costante che variabile) sono molto simili. Le differenze tra i classificatori sono minime, indicando che la regressione logistica, con o senza adattamento del learning rate, riesce a ottenere prestazioni comparabili a quelle del classificatore Bayesiano ottimale.

```
Accuracy Bayesiano: 0.7344
Errore Bayesiano: 0.2656
Accuracy Regressione Logistica (LR Costante): 0.7337
Errore Regressione Logistica (LR Costante): 0.2663
Accuracy Regressione Logistica (LR Variabile): 0.7312
Errore Regressione Logistica (LR Variabile): 0.2688
```

LIST OF FIGURES