UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE ED
ELETTRICA E MATEMATICA APPLICATA

Computer Engineering

Data Analysis

Report prepared during the class session

30/01/2025

Professors:
Prof. Postiglione Fabio
Prof. Matta Vincenzo

Students:
Giovanni Casella
Christian Salvatore De Angelis
Nunzio Del Gaudio

YEAR 2024/2025

Contents

Chapter 1

Itroduction

This report was completed in 4 hours for a Data Analysis exam held on January 30, 2025, at the Department of Computer Engineering of the University of Fisciano.

The purpose of this report is to demonstrate the ability to apply the theoretical and practical knowledge acquired during the course to analyze the provided data, generate a linear regression model, and perform binary classification.

A given prompt was provided, and our goal was to make the best use of the available time to develop both an appropriate code implementation and a report that highlights the key elements of our analysis.

1.1 Assignment

Data Science / Data Analysis — 30 gennaio 2025

Laurea Magistrale in Ingegneria Informatica – Università degli Studi di Salerno

Tempo a dispozione: **4 ore**.

Quesito 1.

Si analizzi il data set `RegressionDSDA250130.csv` che contiene $n = 100$ osservazioni di una variabile dipendente Y e di $p = 25$ regressori X_j $(j = 1, 2, \ldots, p)$, tutti potenzialmente utili alla predizione di Y. A tal fine, si richiede di utilizzare l'ambiente **R**.

1) Determinare i modelli lineari che minimizzano il Mean Squared Error (MSE), individuando i regressori significativi per la predizione di Y e stimando i loro coefficienti β_j, tramite le seguenti strategie:
   i) **best subset selection (BSS)** basata sul Bayesian Information Criterion (BIC),
   ii) **backward stepwise** che utilizza 5**-fold cross-validation**,
   iii) **ridge regression**,
   iv) **LASSO regression**.
2) Valutare l'MSE di test dei modelli lineari individuati al punto precedente e selezionare la strategia di regressione che permette di costruire il modello empirico lineare che minimizza, tra quelle esaminate, l'MSE di test.

Si richiede che il 70% delle osservazioni del data set per la regressione venga utilizzato per il training dei modelli e la scelta dei loro parametri, mentre il test set sia costituito dal restante 30% dei dati forniti.

Quesito 2.

Per la parte da sviluppare al calcolatore, lo studente può utilizzare indifferentemente MATLAB, Python, o entrambi.

Si consideri un problema di classificazione binario descritto da due variabili aleatorie, l'ipotesi $Y \in \{-1, 1\}$ e l'osservazione $X \in \mathbb{R}$.

Caso 1: modello statistico perfettamente noto.

Le due ipotesi sono equiprobabili a priori, vale a dire:

$$\pi(-1) = \mathbb{P}[Y = -1] = \frac{1}{2}, \qquad \pi(+1) = \mathbb{P}[Y = +1] = \frac{1}{2}. \tag{1}$$

La distribuzione condizionata delle feature X dato Y è Gaussiana, con varianza pari a σ^2 e media che dipende dall'ipotesi, specificamente:

$$\ell(x|Y = -1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x+1)^2}{2\sigma^2}\right\}, \qquad \ell(x|Y = +1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-1)^2}{2\sigma^2}\right\}. \tag{2}$$

1) Si calcoli la pmf a posteriori, $p(y|x) = \mathbb{P}[Y = y|X = x]$.
2) Si scelgano poi due valori per σ^2, indicati come σ^2_{easy} ed σ^2_{diff}, che siano rispettivamente rappresentativi di un problema di classificazione "facile" e di uno "difficile". **(Si consiglia di scegliere valori di varianza:** i**) non "estremi", in modo da evitare probabilità di errore troppo piccole o troppo prossime a** $1/2$**;** e ii**) sufficientemente diversi in modo da evidenziare le differenze tra i due scenari).** Si rappresenti graficamente al calcolatore la funzione $p(+1|x)$ al variare di x, per i due valori di varianza scelti. Si commenti il risultato ottenuto, mettendo in relazione la forma delle curve rappresentate e la difficoltà del problema di classificazione.
3) Si valuti empiricamente, attraverso simulazione Monte Carlo, la probabilità di errore del metodo MAP per i due valori di varianza scelti, e si commenti il risultato.

Caso 2: classificazione supervisionata.

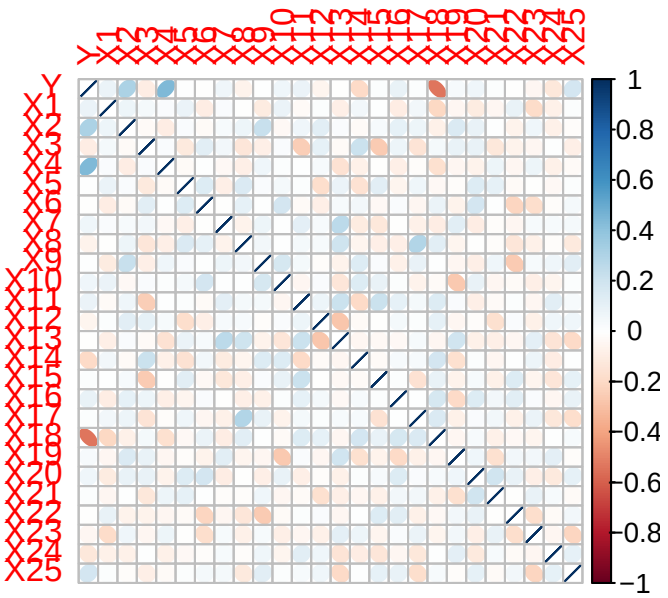
Si generi ora un training set assumendo il modello sopra descritto, per il solo valore di varianza σ^2_{easy}. Lo studente è libero di selezionare un numero di esempi sufficientemente grande da garantire buone prestazioni degli algoritmi di apprendimento da implementare nel seguito.

1) Utilizzare il metodo della regressione logistica per la classificazione binaria, addestrando il sistema con un algoritmo del gradiente stocastico.
2) Utilizzando i parametri stimati al punto precedente, calcolare empiricamente le prestazioni (in termini di probabilitá di errore) del classificatore ottenuto al punto precedente. Confrontare i risultati ottenuti con il caso di modello noto e commentare adeguatamente.

Chapter 2

Regression (Quesito 1)

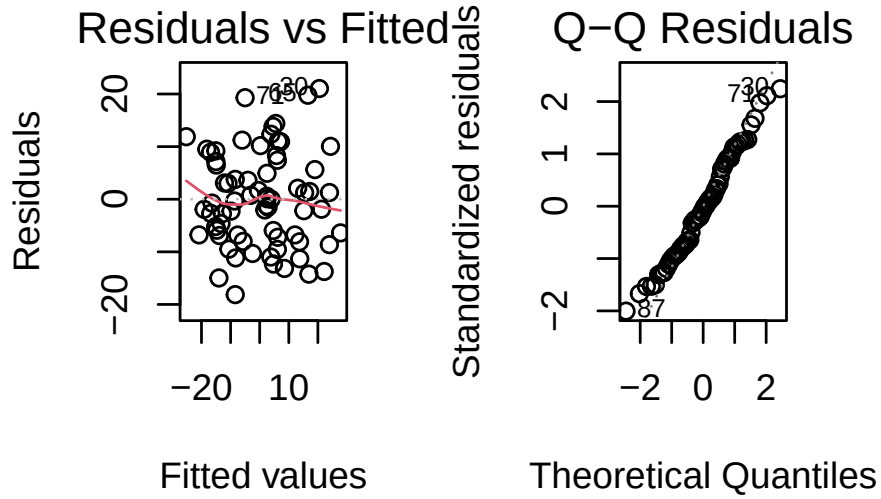
2.1 Datga evaluation

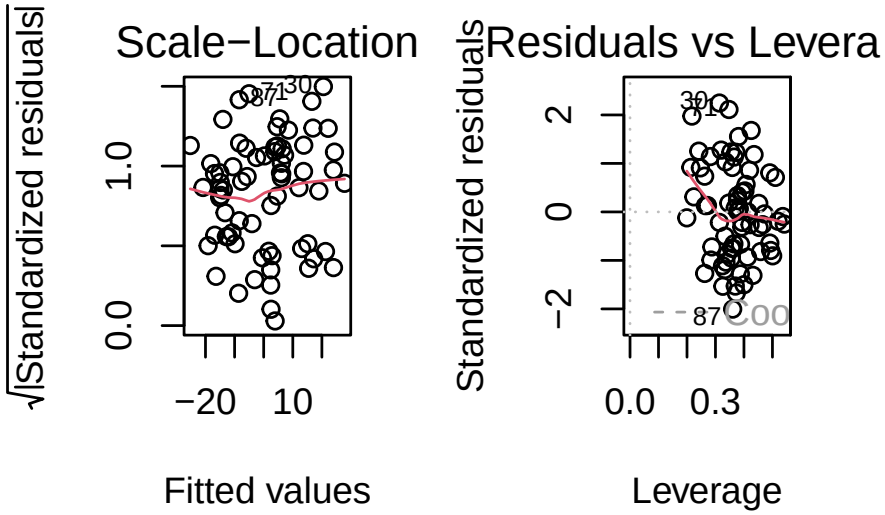


First, we assess a possible linear correlation between the data before proceeding.

We generate a correlation plot for all the variables involved (including Y). There do not appear to be any evident linear correlations between the regressors.

For this reason, we do not conduct further evaluations on this aspect.





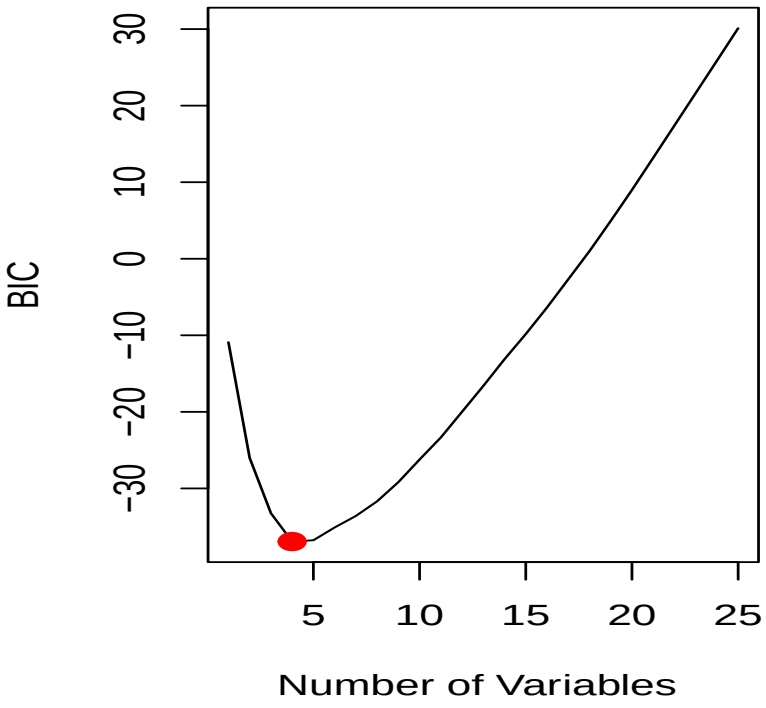
We then evaluated a model composed of all the regressors to make the following considerations:

- There is no collinearity between the regressors, as no VIF values fall between 5 and 10.
- There is no correlation between the error terms.
- The error can be considered Gaussian.
- There are no outliers.
- There are no leverage points.

2.2 Regression techniques

In this section, we will discuss the results of some regression techniques required by the assignment.

2.2.1 Beast subset selection with BIC

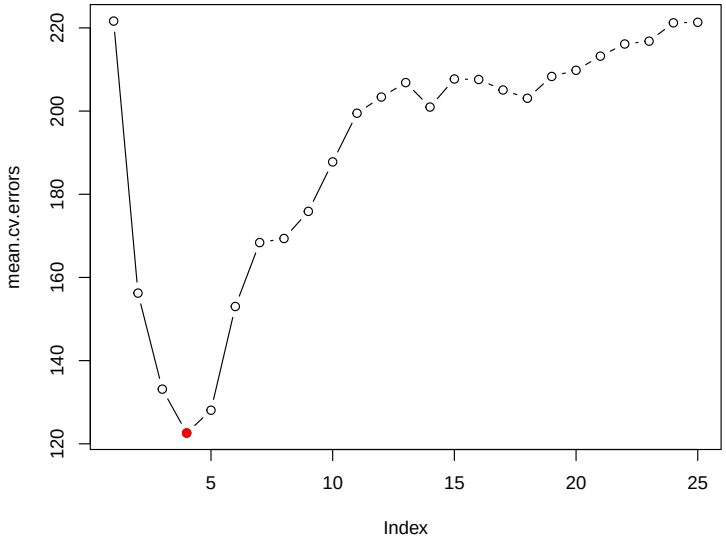


Using the BSS algorithm, the optimal number of regressors is 4, and they are:

```
(Intercept)          X2          X4         X18         X25
 -0.5004706   1.7430057   2.5100572  -2.3708329   1.1741859
```

The One Standard Rule was not considered, as the model with only three regressors has a significantly higher BIC score. Instead, it might be reasonable to retain five regressors, given that the BIC value is very similar to that of the four-regressor model, in order to achieve greater predictive capability.

2.2.2 Backward with cross-validation

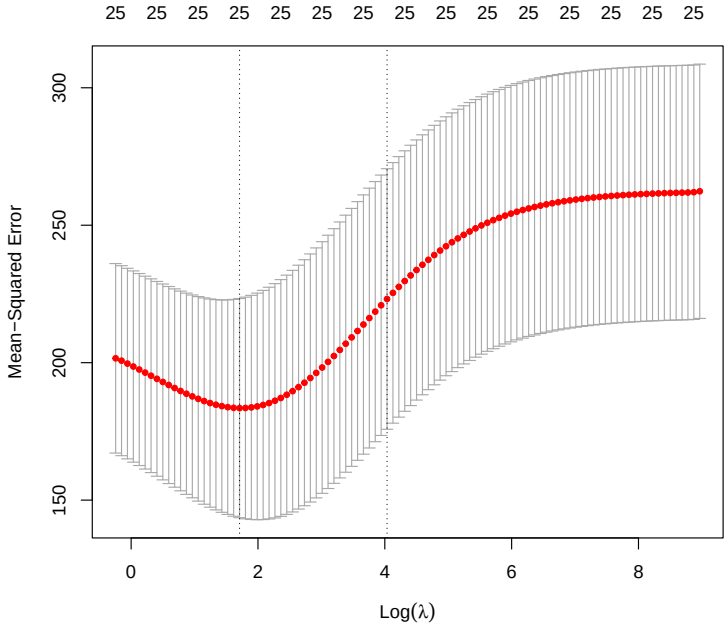



```
(Intercept)          X2          X4         X18         X25
 -0.5004706   1.7430057   2.5100572  -2.3708329   1.1741859
```

We use the cross-validation technique to obtain the best model using the backward method. Since the dataset contains more than a hundred samples, we can also evaluate the presence of the maximum number of regressors, which is 25.

We observe that this technique produces exactly the same model as the Best Subset Selection method. This means that the same model was chosen both to achieve the best trade-off between RSS and the number of regressors and in cross-validation (which should simulate the results on a test set).

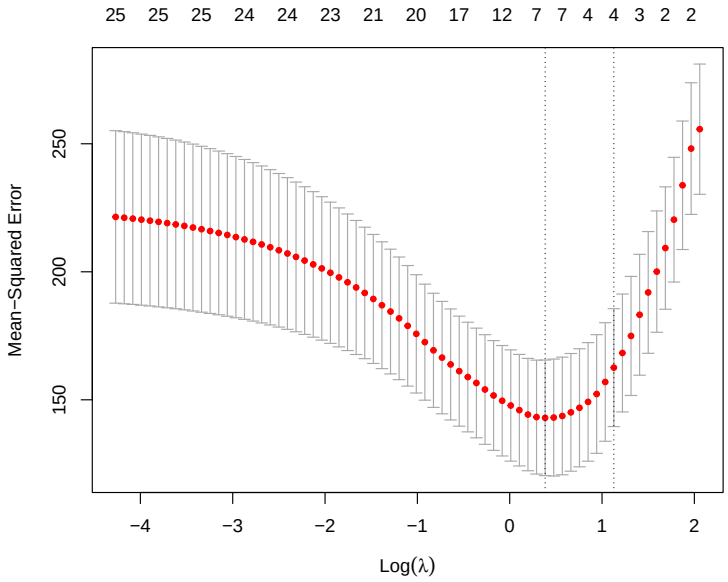
2.2.3 Ridge with cross-validation




```
(Intercept)  0.68373381
X1          -0.17325681
X2           1.10640319
X3          -0.44381225
X4           1.93870132
...
```

The Ridge technique gives the best result for $\lambda = 5.52$. Obviously, Ridge does not perform variable selection. However, we avoid reporting all the regressors of the developed model.

2.2.4 Lasso with cross-validation




```
Lasso:                          Lasso One-standard-rule:
----------------------          -------------------------
(Intercept) -0.1311871          (Intercept)  0.04447872
X2           1.1093006           X2          0.54396966
X3          -0.1419391
X4           2.0903498           X4          1.60303649
X11          0.1313176
X16          0.3241474
X18         -1.9413102           X18        -1.38240976
X25          0.6318643           X25         0.08073831
```

Unlike Ridge, Lasso performs variable selection. Specifically, we considered the model with 7 regressors (with a lower λ and the lowest MSE on validation) and the model with 4 regressors (according to the one-standard-error rule), which allows us to maintain a simpler model.

2.2.5 Our propose

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7767     1.0811  -0.718  0.47424
X2            1.8014     0.3630   4.962 3.09e-06 ***
X4            2.4464     0.4037   6.060 2.80e-08 ***
X9           -0.7123     0.3824  -1.863  0.06563 .
X18          -2.4408     0.3764  -6.484 4.11e-09 ***
X25           1.1817     0.3542   3.337  0.00122 **
```

We also develop a model with 5 regressors, selecting as the 5th regressor the one that, in the Best Subset Selection method, would have slightly increased the BIC parameter.

We observe that the added regressor, although having a p-value low enough to consider it nonzero, has a very weak acceptance strength (i.e., the confidence in considering it different from zero is low).

2.3 Final evaluation

Valutiamo in fine i ristultati sul test set

```
mse_Ridge = 130
mse_lasso = 119
mse_lasso_osr = 151
mse_bss = 119
mse_backward = 119
our_propose = 117
```

Analyzing the MSE values on the test set, it is clear that the model with 4 regressors is the most interesting (the models with the same MSE error also have 4 regressors). The only difference between these models is the value of the coefficients, which, as expected, are smaller for Lasso.

We therefore recommend the Lasso model with 4 regressors as the simplest and most easily interpretable choice. However, we also propose an alternative model with 5 regressors, developed by us, for those who aim to achieve the best possible performance on the test set.

As a final consideration, we believe that the 4-regressor model is the most robust and reliable.

Chapter 3

Classification (Quesito 2 e 3)

3.1 Analytical calculation of the posterior PMF (Quesito 2)

We report only the key points in deriving the posterior probability for the +1 case.

$$P(+1|x) = \frac{P(x) * P(x|+1)}{P(x) * P(x|+1) + P(x) * P(x|-1)} =$$

$$= \frac{\frac{1}{2} * e^{\frac{-(x+1)^2}{2\delta^2}}}{\frac{1}{2} * e^{\frac{-(x-1)^2}{2\delta^2}} + \frac{1}{2} * e^{\frac{-(x+1)^2}{2\delta^2}}} =$$

$$= \frac{1}{1 + e^{\frac{-4x}{2\delta^2}}}$$

3.1.1 Data distribution

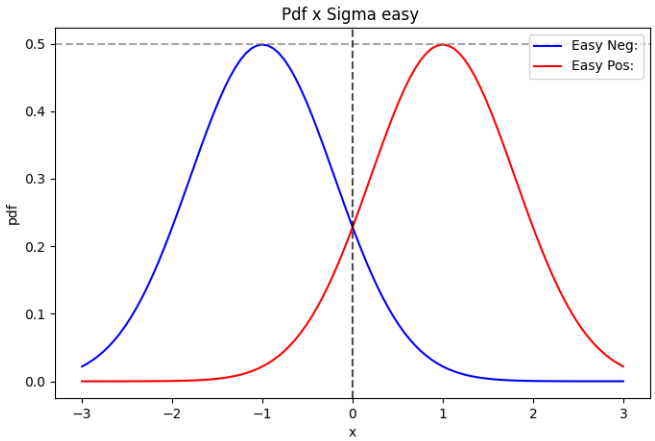


Figure 3.1: Caso facile.

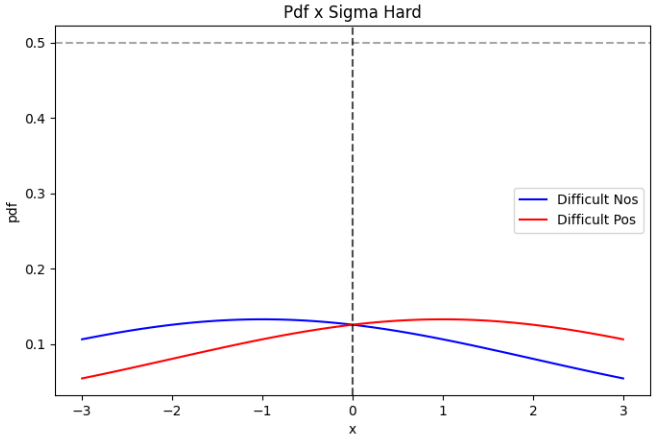


Figure 3.2: Caso difficile.

In the data generation process, as required, we used different variance values, specifically:

- $\sigma^2_{\text{easy}} = 0.8$, since this value prevents the two Gaussians (the feature distributions) from overlapping. As a result, the classification task becomes easier because the distributions are well separated and only intersect at the tails.
- $\sigma^2_{\text{diff}} = 2$, where the two Gaussians completely overlap, making classification significantly more challenging.

3.1.2 Postirior Pmf evaluation

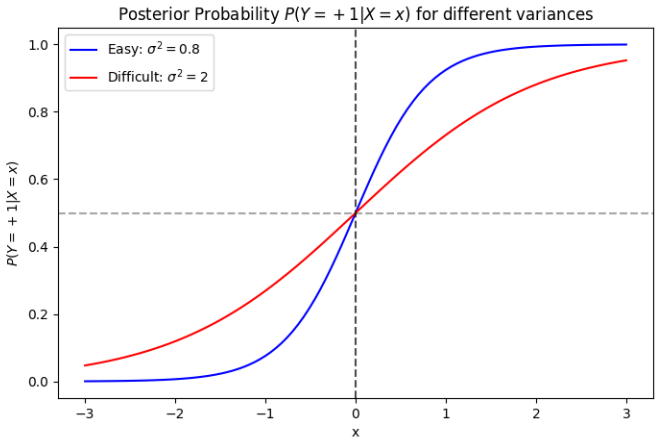


Figure 3.3: Obtained through the simple substitution of values into the posterior formula.

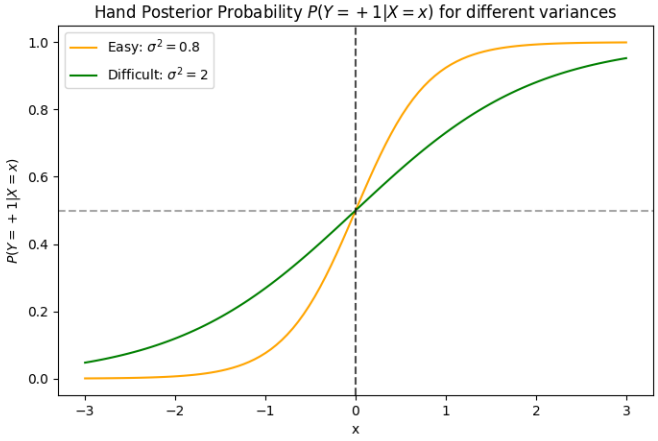


Figure 3.4: Obtained through the analytical formula, computed as indicated in the analytical calculation section.

The probability functions as x varies indicate the probability of assigning the data point x to class $+1$.

We observe that as the variance increases, the function increasingly takes the shape of a straight line, and its intercept grows. For an even higher variance, the function would become flat, making it impossible to determine the class membership unless there are differences in the prior values.

Another consideration is that, since the two Gaussians are equidistant from zero, the posterior probability at $x = 0$ will always be 50%, regardless of the variance value.

3.1.3 Probabilità errore MAP

```
Accuracy MAP EASY 0.866
Accuracy MAP DIFF 0.722
ErrorProb MAP EASY 0.134
ErrorProb MAP DIFF 0.278
```

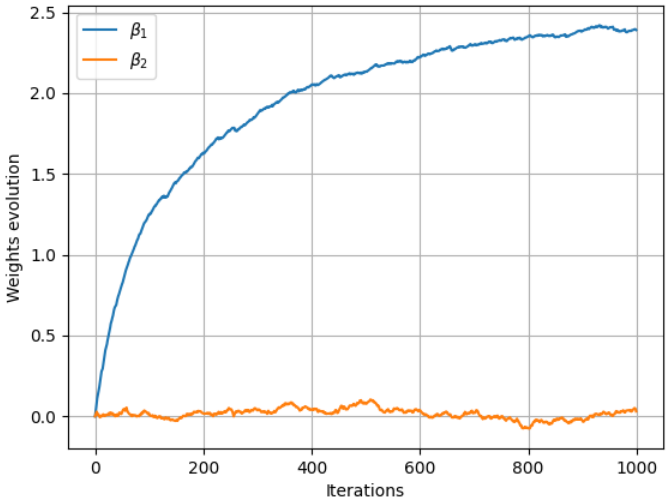
We then evaluate, through a Monte Carlo simulation, the correct classification of the data in both the easy and difficult cases.

Using the MAP method, the accuracy in the easy case is approximately 85%, while in the difficult case, it is 75%. This demonstrates the challenge for the classifier when the variance differs.

The test was conducted on 1000 samples over 5 Monte Carlo iterations. (Each Monte Carlo iteration was performed on a different test set.)

By increasing σ^2_{diff}, we observe that accuracy decreases, and consequently, the probability of error increases.

3.2 Logistic regression (Quesito 3)



These are the parameters obtained through logistic regression with a constant step size.

The stochastic gradient algorithm was executed with 1000 training samples and 10 Monte Carlo iterations.

We observe that only B_1 has a nonzero value, which was expected from the functional form obtained analytically earlier.

3.2.1 Prestazioni Regressione Logistica

```
Accuracy SGD EASY 0.866
ErrorProb SGD EASY 0.134
Accuracy SGD_LR EASY 0.866
ErrorProb SGD_LR EASY 0.134
```

The performance of the model trained using logistic regression is very similar to that obtained using MAP with the known model (both with fixed and variable learning rates).