Práctica 2: Regresión Logística Bayesiana para Predicción de Fallo Cardíaco

Diego Esclarín Fernández

3 de diciembre de 2025

Índice

1. Introducción **2**

2. Metodología **2**
2.1. Regresión Logística Bayesiana 2
2.2. Algoritmo Metropolis-Hastings 2

3. Descripción del Dataset **3**

4. Implementación **3**

5. Resultados **3**
5.1. Interpretación de Coeficientes 4
5.2. Análisis 4

6. Conclusión **4**

1. Introducción

El objetivo de esta práctica es desarrollar un modelo de **Regresión Logística Bayesiana** para predecir la mortalidad en pacientes con insuficiencia cardíaca. A diferencia de la regresión logística clásica, el enfoque bayesiano nos permite estimar no solo los coeficientes del modelo, sino también la *incertidumbre* de cada uno de ellos.

2. Metodología

2.1. Regresión Logística Bayesiana

En la regresión logística, modelamos la probabilidad de que una variable binaria `DEATH_EVENT` sea 1 dado un vector de características. La función de enlace utilizada es la sigmoide:

$$P(y = 1|X, w) = \sigma(X^T w) = \frac{1}{1 + e^{-X^T w}}$$

Donde w son los pesos o coeficientes del modelo.

En la regresión logística Gaussiana, tratamos w como variables aleatorias con una distribución $P(w)$, normalmente Gaussiana. Calculamos la distribución a posteriori $P(w|D)$ con los datos $D = \{(X_i, y_i)\}$, utilizando el Teorema de Bayes:

$$P(w|D) \propto P(D|w)P(w)$$

2.2. Algoritmo Metropolis-Hastings

Como la integral para calcular la posterior exacta no se puede calcular, utilizamos métodos de Monte Carlo Markov Chain (MCMC). Este algoritmo nos permite obtener muestras de la distribución a posteriori siguiendo estos pasos:

1. Inicializar los pesos w aleatoriamente (funciona un poco mejor que inicializarlo a cero como se puede comprobar al ejecutar el experimento).
2. Proponer un nuevo estado w' añadiendo ruido gaussiano: $w' = w + \epsilon$.
3. Calcular la razón de aceptación $\alpha = \text{mín}(1, \frac{P(D|w')}{P(D|w)})$.
4. Aceptar w' con probabilidad α.
5. Repetir el proceso por N iteraciones.

3. Descripción del Dataset

El conjunto de datos utilizado es `fallo_cardiaco.csv`, que contiene registros clínicos de pacientes.

- **Total de muestras:** 299 pacientes.
- **Variable objetivo:** `DEATH_EVENT` (1: Fallecido, 0: Sobrevivió).
- **Características principales:** Edad, fracción de eyección, creatinina sérica, sodio sérico, tiempo de seguimiento, entre otras.

En el proceso se normalizan las variables numéricas para facilitar la convergencia del algoritmo MCMC.

4. Implementación

El proyecto se estructura de la siguiente manera:

- `src/models/bayesian_logistics.py`: Contiene la clase `BayesianLogisticRegression` que implementa el algoritmo Metropolis-Hastings.
- `src/utils/`: Módulos para carga de datos, preprocesamiento y división del dataset.
- `train.py`: Script principal que orquesta la carga y entrenamiento del modelo.
- `test.py`: Script que evalúa el modelo entrenado.

5. Resultados

El modelo fue entrenado con 6000 muestras y un periodo de *burn-in* para asegurar la convergencia.

5.1. Interpretación de Coeficientes

A continuación se presentan los coeficientes medios aprendidos por el modelo y su desviación estándar (incertidumbre):

Variable	Influencia (Media)	Incertidumbre (Std)
Age	0.56	0.23
Anaemia	-0.08	0.17
Creatinine Phosphokinase	0.16	0.18
Diabetes	0.1	0.22
Ejection Fraction	-0.68	0.18
High Blood Pressure	-0.05	0.18
Platelets	-0.01	0.18
Serum Creatinine	1.17	0.47
Serum Sodium	-0.12	0.16
Sex	-0.31	0.16
Smoking	0.04	0.15
Time	-1.36	0.19

Cuadro 1: Resumen de los parámetros del modelo posterior.

5.2. Análisis

- **Factores de Riesgo:** La **Creatinina Sérica** (1.17) y la **Edad** (0.56) son los factores que más aumentan la probabilidad de muerte. Sin embargo, la creatinina tiene una alta incertidumbre (0.47), lo que sugiere variabilidad en su impacto entre pacientes.
- **Factores Protectores:** El **Tiempo** de seguimiento (-1.36) y la **Fracción de Eyección** (-0.68) tienen coeficientes negativos fuertes, indicando que valores altos en estas variables reducen significativamente el riesgo.
- **Variables Irrelevantes:** Variables como *Platelets* y *Smoking* tienen coeficientes cercanos a cero, sugiriendo poca influencia en este modelo específico.

6. Conclusión

El modelo bayesiano ha permitido identificar los principales factores de riesgo asociados al fallo cardíaco. La capacidad de estimar la incertidumbre es una ventaja clave, permitiéndonos saber no solo qué variables son importantes, sino también qué tan seguros estamos de esa importancia. Aunque algunos resultados como que si fumar es un factor irrelevante para este análisis sea contradictorio con el conocimiento general.

Con estos datos el modelo tiene un accuracy del 83.33 %.