Estimación de un modelo de aprendizaje con regresión logística utilizando métodos de optimizacion numérica

1 Introducción

Tomando como referencia el ejercicio planteado en clase, el cual toma como referencia el artículo publicado por Colubri et. al (2019), se busca entrenar un modelo de regresión logística que permite pronosticar la supervivencia a enfermedad por virus del ébola. Para tal propósito, se utilizan datos recolectados por el Cuerpo Internacional de Medicina (IMC) durante 2014 y 2016 en Liberia y Sierra Leona. El fin de este trabajo es explorar los resultados de pronóstico y desempeño computacional derivados de la utilización de diferentes métodos de optimización numérica aplicados a machine learning, y de la implementación de métodos de cómputo en paralelo durante el proceso de estimación.

2 Objetivos

El objetivo principal de este trabajo es evaluar el rendimiento de los métodos de optimización mencionados en la sección 5. Dado lo anterior, definimos rendimiento en términos de métricas de desempeño del modelo, y también en métricas de desempeño durante el proceso de computo. A continuación, se describe el planteamiento del problema a resolver.

2.1 Problema a resolver

El método de *regresión logística* asume que $Pr[y_i|x_i, \beta] \sim Bernoulli(\mu_i)$, con:

$$\mu_i = \sigma(\beta^T x_i)$$
$$\sigma(z) = (1 + exp(-z))^{-1}$$
$$\beta \in \mathbb{R}^p$$

El objetivo es encontrar el modelo $\hat{\beta} \in \mathbb{R}^p$ que mejor se ajuste al conjunto de datos. Para estimar $\hat{\beta}$, implementaremos y compararemos tres métodos: método de Newton, método Broyden, Fletcher, Goldfarb y Shanno(BFGS) y el método del descenso de gradiente estocástico (SGD, por sus siglas en inglés).

Dado lo anterior, se resuelve el problema de manera iterativa por medio de la resolución de los siguientes objetivos intermedios:

- Modelar $Pr[y|x, \hat{\beta}]$
- Predecir la etiqueta $\hat{y} \in 0, 1$ de un nuevo dato x por medio de:

$$\hat{y} = \begin{cases} 0, & \text{si } \sigma(\hat{\beta}^T x) \geq 0.5 \\ 1, & \text{si } \sigma(\hat{\beta}^T x) < 0.5 \end{cases} \tag{1}$$

- Computar la función de pérdida correspondiente a la *log-verosimilitud negativa* (y el riesgo empírico en SGD):

$$F(\beta) := LVN(\beta) = -\sum_{i=1}^{m} [y_i log\mu_i + (1 - y_i)log(1 - \mu_i)] \tag{2}$$

- Comparar tiempo de convergencia entre los tres métodos y las soluciones alanzadas en términos de las métricas de desempeño del modelo.

3 Revisión de Literatura

Como se mencionó al principio, el presente trabajo se apoya principalmente en el artículo *Machine Learning Models from 2014-2016 Ebola Outbreak-Data-harmonization Challenges, Validation, Strategies and mHealth Applications*, cuyo objetivo principal fue diseñar una app que, dado los síntomas de un paciente con ébola, calcula la probabilidad de muerte del paciente y recomienda un tratamiento personalizado. Lo anterior, es resultado de la implementación de diferentes modelos de regresión logística, mediante los cuáles se identificaron los principales factores asociados a la muerte de una persona con ébola (case fatality rate, CFR).

3.1 Internal Medical Corps (IMC): Conjunto de datos de Entrenamiento

Considerando la situación sanitaria del 2014-2016, la app fue focalizada en regiones con mayor afectación y menor posibilidad de acceso a atención médica. Así, los modelos se entrenaron con datos provenientes de cinco unidades de tratamiento del ébola (ETU, por sus siglas en inglés), localizadas en Sierra Leona y Liberia. La base constaba con 470 observaciones (correspondientes a los pacientes atendidos), con variables como temperatura, escala de bienestar, time to presentantion (TTP), triaje, entre otras variables; además, a partir de la reacción en cadena de polimerasa se calculó el cycle threshold (CT).

3.2 Análisis bivariado y Exploratorio

Con estos datos se procedió con un análisis bivariado y exploratorio entre la variable dependiente (muerte o no muerte)y el resto de las variables, identificando la asociación entre cada uno de los posibles regresores y la variable dependiente. Para evaluar la asociación entre las variables numéricas y la variable dependiente, se realizaron *point biseral correlation tests*; para evaluar la asociación entre dos variables categóricas, se realizaron pruebas chi cuadradas con corrección de Yates.

El resultado de este primer análisis fue que más de 50 por ciento de los pacientes que murieron por ébola reportaron síntomas de pérdida de apetito, fiebre, astenia, dolor musculoso esquelético, dolor de cabeza y diarrea. Además, se identificó prevalencia de triajes notablemente diferente entre los resultados (mortalidad y no no mortalidad de los pacientes). Finalmente, observando el *p-value* se identificó que pocas variables se asocian significativamente con el resultado del paciente: CT, edad, ictericia (jaundice) –con un p-value menor a 0.05; conjuntivitis, confusión, disnea, dolor de cabeza y sangrado – con un p-value menor a 0.15.

3.3 Imputaciones Múltiples

Un problema suscitado al trabajar con estos datos, fue la cantidad de valores faltantes (22 por ciento respecto del total de potenciales predictores). Con base en lo anterior, se tomaron las

siguientes acciones: se eliminaron los predictores potenciales que tuvieran más de 50 por ciento de valores faltantes; se evaluó la idoneidad de imputar los valores faltantes mediante la prueba Little's MCAR, cuya hipótesis nula establece que los datos faltantes se debe totalmente al azar y se implementa probando si las medias de las variables de interés, a través de distintos grupos no cambian. Resultado del test, se aceptó que los datos faltantes se debían completamente al azar y se realizó una imputación múltiple con la función aregImpute del paquete de R Hmisc. En resumen, la imutación consistió en la generación de: una distribución predictiva bayesiana a partir de los datos conocidos y un número N de conjuntos de datos imputados. Cada valor faltante en la i-ésima imputación se predijo a partir de un modelo aditivio ajustado en una muestra de bootstrap con reemplazo de los datos originales.

3.4 Selección de Variables

Con el fin de contar con un modelo parsimonioso, los autores ejecutaron un Elastic Net Regularization, misma que combina las normas $L1$ y $L2$ de las regresiones de Lasso y Ridge, para la selección del subconjunto de características relevantes y no redundantes. Los criterios de selección de los regresores candidatos no redudantes y relevantes fueron los siguientes: se mantuvieron las variables que tuvieron un coeficiente positivo en al menos 50 por ciento de las regresiones penalizadas; se eliminaron las variables con p-value bajo y se incorporaron las variables que tuvieron una asociación débil en el análisis bivariado. El resultado fue un conjunto de variables que se usaron en un modelo denominado parsimonioso.

Además, las variables seleccionadas fueron evaluadas por sesgo; luego, se aplicaron splines cúbicos restringidos para modelar las relaciones no lineales entre CFR y edad y CFR y temperatura corporal, donde los valores reales de edad y temperatura fueron la entrada del procedimiento de ajuste del spline, tal como se implementó en el paquete Hmisc.

3.5 Modelos de Regresión Logística: Modelo Parsimonioso y Modelos Alternativos

Derivado del análisis y procedimiento anterior, se implementaron varios modelos: el modelo parsimonioso – edad del paciente, edad del paciente al cubo, temperatura corporal, temperatura corporal al cubo, amarillamiento, sangrado, disnea, disfagia, TTP y CT x TTP– y tres modelos adicionales que se emplearon cuando los datos de triaje eran limitados: parsimonioso sin temperatura –idéntico al parsimonioso pero con eliminación de la temperatura corporal–, sólo clínico –ictericia, sangrado, disnea, disfagia, debilidad y diarrea– y el modelo mínimo –TC y edad. Cada modelo final se obtuvo ajustando N copias del modelo, en cada conjunto de datos imputados, y promediando esas copias en un solo modelo usando 'fit.mult.impute' de Hmisc.

Una predicción se clasificó como muerte cuando la puntuación del modelo superó el umbral de 0.5. Los intervalos de confianza (IC) de todas las estimaciones de rendimiento se calcularon utilizando la transformación de Fisher. Las razones de probabilidad (OR) de la regresión logística se convirtieron en razones de riesgo (RR).

3.6 Resultados

El modelo parsimonioso representó de manera adecuada las probabilidades reales de muerte, con cierta subestimación para pacientes de bajo riesgo de muerte.

Los modelos mínimos y solo clínicos resultaron menos calibrados, con el primero subestimando las probabilidades reales tanto en el lado de bajo y alto riesgo, y el segundo sobreestimando las probabilidades reales para una amplia gama de riesgos por encima del 60 por ciento.

Los índices de validación de todos modelos, junto con sus intervalos de confianza (al 95 por ciento), se superponen. La especificidad y R2 fue mayor en el modelo parsimoniosos. Mientras que la sensibilidad fue ligeramente mayor en el modelo mínimo. El modelo clínico muestró un rendimiento consistentemente menor.

Para el caso del modelo parsimonioso, la clasificación de todas las variables por su importancia, medida por el estadístico Wald $\tilde{\chi}^2$ – prueba estadística paramétrica que se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra–, están dadas en el siguiente orden: CT, edad del paciente, ictericia, hemorragia, temperatura corporal y la interacción $CT - TTP$.

En términos de los odds ratios del modelo parsimonioso, se concluyó que la presentación de ictericia o sangrado se asocia con más del doble de las probabilidades de muerte, aunque su prevalencia es baja (al 5 por ciento). Finalmente, al transformar los riesgos de probabilidad en razones de riesgo, se identificó una estimación del 4 por ciento de aumento en el riesgo asociado con la aparición de esos síntomas.

3.7 Validación externa: KGH y GOAL

Los resultados fueron validados con dos bases de datos –provenientes del Kenema Government Hospital (KGH) y GOAL–; el primero constaba con 106 casos con ébola y una tasa de fatalidad (CFR, por sus siglas en inglés) de 73 por ciento; el segundo constaba de 158 casos con ébola, con una tasa de fatalidad de 60 por ciento.

Al igual que la base de datos de entrenamiento, ambas bases de datos empleadas durante la validación, contaban con registros incompletos, razón por la cuál se les aplicaron también imputaciones múltiples.

En general, los resultados de la validación fueron los siguientes: todos los modelos, con excepción del modelo clínico, exhibieron un rendimiento consistente en términos del área bajo la curva (AUC) y precisión general, con diferencias análogas en términos de especificidad y sensibilidad. El modelo parsiomonioso presentó mayor especificidad, mientras que el modelo mínimo fue más sensible. El impacto de la imputación fue intrascendente.

4 Conjunto de Datos

Por la restricciones de uso de la base de datos de *entrenamiento* original (además de una serie de procedimientos díficiles de seguir como protocolos de investigación y evidencia de aprobación por nuestro protocolo de investigación por parte de un Comité de ética independiente), optamos por trabajar con una de las dos bases de datos que los autores emplearon para *validar* sus modelos: KGH. La base de datos en mención, como se mencionó en la sección anterior, consta de 106 casos positivos de pacientes con ébola y un CFR global de 73 por ciento. Originalmente, previo al tratamiento de los datos, la base tenía únicamente 44 registros de triaje, 58 registros de carga viral, con un total de 78 valores faltantes en todo el data set.

Para harmonizar los datos, los autores transformaron la carga viral en CT, conforme con la curva estándar qPCR:

$$log_{(carga\ viral)} = m * CT + c_0$$

Los resultados de la transformación no dependieron del origen geográfico de los datos IMC. Sin embargo, el origen geográfico tuvo un impacto significativo ($p < 0.0001$) en la distribución de CT a través de los sitios.

Tipo	Nombre	Descripción
Variables Numéricas	CT	El cycle threshold (CT) es una variable que se calcula a partir de una relación médica bien conocida (qPCR) y la carga viral (una expresión númerica de la cantidad de virus dado un volúmen de fluido que normalmente se correlaciona con la severidad de una infección viral activa).
	TEMP	Temperatura corporal del paciente. Toma valores de 36 a 39.9
	AGE	Edad del paciente. Toma valores de 1 a 73.
Variables Categóricas	HEADCH	Presencia o no dolores de cabeza. Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no dolores de cabeza.
	BLEED	Presencia o no de sangrado. Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no sangrado.
	DIARR	Presencia o no de diarrea. Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no diarrea.
	VOMIT	Dificultad para comer, conocido como disfagia, término técnico para describir el síntoma consistente en dificultad para la deglución (problemas para tragar). Esta dificultad suele ir acompañada de dolores, a veces lancinantes (disfagia dolorosa u odinofagia . Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no de disfacia.
	PABD	Presencia o no de PADB.
	WEAK	Presencia o no de debilidad o fatiga general.Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no debilidad.
	JAUN	Condición en la cuál la piel, los ojos y los miembros mucosos que vuelven amarillos debido a altos niveles de bilirubina. Toma valores valores 1 o 0, dependiendo de si el paciente presenta o no ictericia.
	OUT	Muerte o no muerte del paciente. Toma valores 1 o 0. Dependiendo de si el paciente muere o no muere.

Table 1: Variables del Dataset Kenema Government Hospital (KGH): Versión Imputada 50.

Para corregir dicho sesgo analítico y garantizar que los resultados fueran internamente consistentes en cada sitio y, por lo tanto, comparables entre los sitios, se normalizaron los valores de CT de cada sitio, mediante el escalado de características (restando la media y dividiendo por la desviación estándar).

Como las diferencias en CT también podrían provocar diferencias en los comportamientos de búsqueda de atención, entonces se normalizaron los valores de la CT dentro de los sitios también contribuiría a reducir el efecto de este posible factor de confusión -una variable o factor que distorsiona la medida de asociación entre otras dos variables.

Nosotros, para fines del presente trabajo, empleamos una de las versiones imputadas de esta base de datos, dispuesta en el siguiente sitio: ebola-imc-public, misma que cuenta con 11 variables.

Así, para evaluar el rendimiento de los métodos de optimización mencionados en la sección 5, realizaremos dos regresiones logísticas: el modelo de regresión logística clínicas (ictericia, sangrado, disfagia, debilidad y diarrea), junto con el CT y el modelo de regresión logística simple (CT y edad como regresores). Esto principalmente se debe al limitado insumo inicial (la base de datos dispuesta al público), mismo que contiene un sub-conjunto reducido de las variables originales empleadas por los autores para el entrenamiento del modelo.

5 Implementación

La implementación de este ejercicio se realizará por medio de código escrito en lenguaje Python. En particular, con el fin de minimizar la función de pérdida de log-verosimilitud negativa (y el riesgo empírico en SGD), se incluirán módulos propios para resolver el problema. Los pasos a seguir son los siguientes:

1. Implementar *grad_F* y hess_F en Python.
2. Implementar el método de máximo descenso para minimizar $F(\beta)$. Elegimos un β^0 aleatorio y una tolerancia de $\epsilon = 10^{-8}$.
3. Implementar un clasificador de regresión logística para obtener $\hat{y}$.
4. Implementar el método de Newton para minimizar $F(\beta)$
5. Implementar el método BFGS
6. Implementar el método SGD
7. Implementar variabilidad en tasa de aprendizaje (condición de Armijo).
8. Paralelizar procesos; por ejemplo: resolver en paralelo la dirección del descenso.
9. Dockerizar ambiente
10. Comparar tiempos de ejecución
11. Unittest

Con el fin de generar un entorno aislado que permita evaluar el desempeño del proceso de entrenamiento en término de recursos computacionales, se utilizará una instancia EC2 de Amazon Web Services (AWS). Adicionalmente, para tener un control del entorno virtual asociado, se creará una imagen de docker asociada al repositorio donde se encontrarán todos los códigos de este proyecto.

6 Algoritmos

6.1 Método de Newton

1. Inicializar β^0
2. Mientras "no converge"
   (a) Resolver d^k en $(\nabla^2 F(\beta^k))d^k = -\nabla F(\beta^k)$
   (b) Utilizar Armijo para encontrar una tasa de aprendizaje α^k
   (c) $\beta^{k+1} \leftarrow \beta^k + \alpha^k d^k$

6.2 Método BFGS

1. Aproximar la inversa $\nabla^2 F(\beta^k)$ por medio de una función de rango-2 dada por:

$$\begin{aligned} H^{k+1} &= H^k + \frac{w^k (w^k)^T}{(w^k)^T z^k} - \frac{H^k z^k (H^k z^k)^T}{(z^k)^T H^k z^k} \\ z^k &= \beta^{k+1} - \beta^k \\ w^k &= \nabla F(\beta^(k+1) - \nabla F(\beta^k) \end{aligned} \tag{3}$$

2. Reescribir $F(\beta)$ como función de riesgo empírico.

3. Derivar $\nabla l(\beta^T x_i, y_i)$.

6.3 Método SGD

1. Inicializar β^0

2. Mientras "no converge"

   (a) Reordenar los datos D de manera aleatoria (para garantizar $\mathbb{E}[\nabla l(\beta^T x, y)] = \nabla F(\beta)$

   (b) Para cada $(x_i, y_i) \in D, \beta^{k+1} \leftarrow \beta^k - \eta \nabla l(\beta^T x_i, y_i)$

Implementar el método de gradiente estocástico con minilotes de cardinalidad $q \in [m]$:

$$\beta \leftarrow \beta - \eta \frac{1}{|S|} \sum_{j \in S} \nabla l(\beta^T x_i, y_i),$$

para minilote $S \sim S \subset D : |S| = q$.

7 Resultados

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:




import pandas as pd

df_proc = df_raw
df_proc['INTER_AGE'] = "NA"
print()

1/10*a
df_proc.dtypes



def grad_cost_func(X,y,beta):
    mu=calcula_mu(X,beta)
    return np.matmul(np.transpose(X),mu-y)
```

```
23
24 def mini_lotes(X,y,q=10):
25     cols=X.shape[1]
26     data=np.hstack((X,y[:,None]))
27     np.random.shuffle(data)
28     data=data[0:q]
29     X=data[:,0:cols]
30     y=data[:,cols]
31     return X,y
```

8 Conclusiones

9 Referencias

- Colubri, A., Hartley, M. A., Siakor, M., Wolfman, V., Felix, A., Sesay, T., ... Sabeti, P. C. (2019). Machine-learning Prognostic Models from the 2014–16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications. EClinicalMedicine, 11, 54-6º4.
- Nocedal, J., Wright, S. (2006). Numerical optimization. Springer Science Business Media.

Appendix A Funciones de Python

Appendix B Código de Python