TALLER IA REGRESIÓN LOGÍSTICA

Análisis de una regresión logística implementada en Orange3

Autor 1: juan pablo aristizabal
Autor 2: Juan David Gallego Rangel
Autor 3 : Angelo Gutierrez Correa:

Universidad Tecnológica de Pereira, Risaralda, Colombia
juanpablo.aritizabal@utp.edu.co
fbdavid37@utp.edu.co
angutierrez@utp.edu.co

***Resumen*—En este taller se pretende realizar una regresión logística donde tomamos los valores del archivos Songs.csv, donde la regresión logística nos permite medir las relación entre la variable dependiente, la afirmación que se desea predecir, con una o más variables independientes,el conjunto de características disponibles para el modelo. Para ello utiliza una función logística que determina la probabilidad de la variable dependiente.**
Lo que se busca en estos problemas es una clasificación, por lo que la probabilidad se ha de traducir en valores binarios. Para ellos se utiliza un valor que denominamos umbral. Los valores de probabilidad por encima del valor umbral la afirmación es cierta y por debajo es falsa. Generalmente este valor es 0,5, aunque se puede aumentar o reducir para gestionar el número de falsos positivos o falsos negativos.

***Palabras clave*— Regresión Logística, algoritmo, función logística, predicciones, modelo, variable dependiente, variable independiente, intervalo de confianza**

Abstract*— *In this workshop we intend to perform a logistic regression where we take the values from the Songs.csv files, where the logistic regression allows us to measure the relationship between the dependent variable, the statement to be predicted, with one or more independent variables, the set of features available for the model. To do this, it uses a logistic function that determines the probability of the dependent variable.
What you are looking for in these problems is a classification, so the probability has to be translated into binary values. For them, a value that we call threshold is used. For probability values above the threshold value the statement is true and below it is false. Generally this value is 0.5, although it can be increased or decreased to handle the number of false positives or false negatives

Keywords*— *Logistic Regression, algorithm, logistic function, predictions, model, dependent variable, independent variable, confidence interval

INTRODUCCIÓN

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan como una función de variables explicativas, utilizando una función logística.

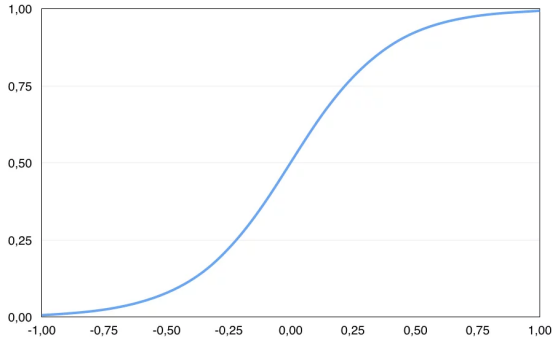
La regresión logística es una técnica de aprendizaje automático que proviene del campo de la estadística. A pesar de su nombre no es un algoritmo para aplicar en problemas de regresión, en los que se busca un valor continuo, sino que es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1. Por ejemplo, un problema de clasificación es identificar si una operación dada es fraudulenta o no. Asociándolo a una etiqueta “fraude” a unos registros y “no fraude” a otros. Simplificando mucho es identificar si al realizar una afirmación sobre registro esta es cierta o no.

FORMULACIÓN MATEMÁTICA DE LA FUNCIÓN LOGÍSTICA

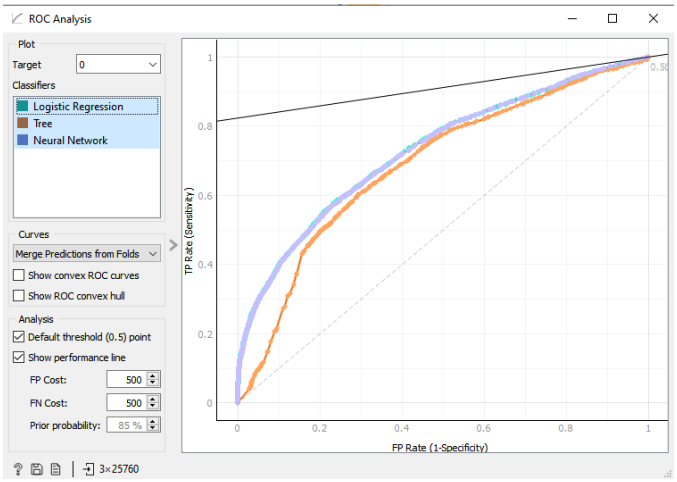
A la función que relaciona la variable dependiente con las independientes también se le llama función sigmoidea. La función sigmoidea es una curva en forma de S que puede tomar cualquier valor entre 0 y 1, pero nunca valores fuera de estos límites. La ecuación que define la función sigmoidea es

$$f(x) = \frac{1}{1 + e^{-x}}$$

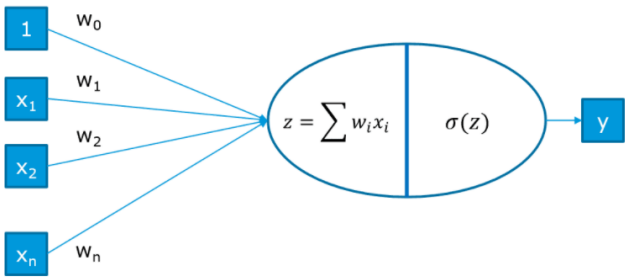
donde X es un número real. En la ecuación se puede ver que cuando X tiene a menos infinito el cociente tiende a cero. Por otro lado, cuando X tiende a infinito el cociente tiende a la unidad. En la siguiente figura se muestra una representación gráfica de la función logística (función sigmoide).



Fecha de Recepción: 22 de Julio del -2021



- Podemos representar lo que hace la regresión logística en la siguiente figura:



VENTAJAS DE LA REGRESIÓN LOGÍSTICA

La regresión logística es una técnica muy empleada por los científicos de datos debido a su eficacia y simplicidad. No es necesario disponer de grandes recursos computacionales, tanto en entrenamiento como en ejecución. Además, los resultados son altamente interpretables. Siendo esta una de sus principales ventajas respecto a otras técnicas. El peso de cada una de las características determina la importancia que tiene en la decisión final. Por lo tanto, se puede afirmar que el modelo ha tomado una decisión u otra en base a la existencia de una u otra característica en el registro. Lo que en muchas aplicaciones es altamente deseado además del modelo en sí. El funcionamiento de la regresión logística, al igual que la regresión lineal, es mejor cuando se utilizan atributos relacionados con la de salida. Eliminado aquellos que no lo están. También es importante eliminar las características que muestran una gran multicolinealidad entre sí. Por lo que la selección de las características previas al entrenamiento del modelo es clave. Siendo aplicables las técnicas de ingeniería de características también utilizadas en la regresión lineal.

DESVENTAJAS DE LA REGRESIÓN LOGÍSTICA

En cuanto a sus desventajas se encuentra la imposibilidad de resolver directamente problemas no lineales. Esto es así porque la expresión que toma la decisión es lineal. Por ejemplo, en el caso de que la probabilidad de una clase se reduzca inicialmente con una característica y posteriormente suba no puede ser registrado con un modelo logístico directamente. Siendo necesario transformar esta característica previamente para que el modelo pueda registrar este comportamiento no lineal. En estos casos es mejor utilizar otros modelos como los árboles de decisión.

Una cuestión importante es que la variable objetivo esta ha de ser linealmente separable. En caso contrario el modelo de regresión logística no clasificará correctamente. Es decir, en los datos han de existir dos "regiones" con una frontera lineal.

Otra desventaja es la dependencia que muestra en las características. Ya que no es una herramienta útil para identificar las características más adecuadas. Siendo necesario identificar estas mediante otros métodos. Finalmente, la regresión logística tampoco es uno de los algoritmos más potentes que existen. Pudiendo ser superado fácilmente por otros más complejos.

APLICACIONES DEL MODELO DE REGRESIÓN LINEAL

Su utilización en la predicción es el uso más frecuente y extendido, enmarcado en los diferentes tipos de estudios, ya sean típicamente prospectivos con finalidad pronóstica (epidemiología clínica), estudios prospectivos con finalidad analítica, estudios caso-control y en los ensayos clínicos. Quisiéramos en este punto resaltar que la RL es un instrumento muy útil para facilitar el tratamiento cuantitativo de los datos pero no podemos aislarlo del diseño del estudio, so pena de cometer errores que nos conducirán a conclusiones erróneas.

CONDICIONES PARA LA REGRESIÓN LOGÍSTICA

Para que un modelo de regresión logística, y las conclusiones derivadas de él, sean completamente válidas, se deben verificar que se cumplen las asunciones sobre las que se basa su desarrollo matemático. En la práctica, rara vez se cumplen, o se puede demostrar que se cumplen todas, sin embargo esto no significa que el modelo no sea útil. Lo importante es ser consciente de ellas y del impacto que esto tiene en las conclusiones que se extraen del modelo.

CONCEPTO DE INTERACCIÓN

Un concepto importante al construir un modelo de regresión es que pueden introducirse términos independientes únicos (una sola variable, por ejemplo efecto del tabaco) y además las interacciones entre variables de cualquier orden (efecto del tabaco según género), si se considera que pueden ser de interés o afectar a los resultados.

Al introducir los términos de interacción en un modelo de regresión es importante para la correcta estimación del modelo respetar un orden jerárquico, es decir siempre que se introduzca un término de interacción de orden superior (x•y•z), deben introducirse en el modelo los términos de interacción de orden inferior (x•y, x•z, y•z) y por supuesto los términos independientes de las variables que participan en la interacción (x, y, z).

ASPECTOS A TENER EN CUENTA PARA EL USO DE LA REGRESIÓN LOGÍSTICA

Tamaño de muestra y número de variables independientes. Una de las ventajas de la regresión logística es que permite el uso de múltiples variables con relativamente pocos casos, sin embargo, hay que tener en cuenta algunas precauciones. Se ha sugerido que el número de sujetos para poder usar esta técnica estadística sin problemas debe ser superior a 10 (k+1) donde k es el número de variables explicativas; por tanto, si se introducen interacciones o variables independientes, el número de elementos en la muestra debe aumentar. Además se ha

sugerido que si una de las variables dicotómicas (en especial si es la de respuesta) no tiene al menos 10 casos en cada uno de sus 2 valores posibles, entonces las estimaciones no son confiables. En cuanto al número de variables independientes, la inclusión de un gran número de ellas en el modelo puede indicar que no se ha reflexionado suficientemente sobre el problema.

Es necesario tener en cuenta el efecto sobre el riesgo de que ocurra el evento, de los cambios de las variables explicativas cuando son cuantitativas (continuas), en ocasiones es necesario categorizarlas, ya que los cambios que se producen de una unidad a otra pueden resultar intrascendentes o no ser constantes a lo largo del rango de valores de la variable.

Cuando algunas de las variables independientes analizadas están altamente correlacionadas, los resultados que se obtienen pueden no ser satisfactorios, por esta razón debe realizarse un análisis previo univariado entre las distintas variables explicativas.9
Para que la regresión logística tenga un sentido claro, tiene que existir una relación monótona entre las variables explicativas y la de respuesta, esto significa que el aumento de las unas se acompañe del aumento o la disminución aproximadamente constante de la otra, para todo el rango de valores estudiados.

La regresión logística es un tipo de modelo lineal.

VALIDACIÓN DEL MODELO

Una vez seleccionado el mejor modelo que se puede crear con los datos disponibles, se tiene que comprobar su capacidad prediciendo nuevas observaciones que no se hayan empleado para entrenarlo, de este modo se verifica si el modelo se puede generalizar. Una estrategia comúnmente empleada es dividir aleatoriamente los datos en dos grupos, ajustar el modelo con el primer grupo y estimar la precisión de las predicciones con el segundo.
El tamaño adecuado de las particiones depende en gran medida de la cantidad de datos disponibles y la seguridad que se necesite en la estimación del error, 80% 20% suele dar buenos resultados

TIPOS DE REGRESIÓN LOGÍSTICA

- Regresión Logística Binaria: la variable objetivo tiene solo dos resultados posibles, Llueve o NO Llueve, Sube o Baja.
- Regresión Logística Multinomial: la variable objetivo tiene tres o más categorías nominales, como predecir el tipo de vino.
- Regresión Logística Ordinal: la variable objetivo tiene tres o más categorías ordinales, como clasificar un restaurante o un producto del 1 al 5

ENTRENAMIENTO DE UNA REGRESIÓN LOGÍSTICA

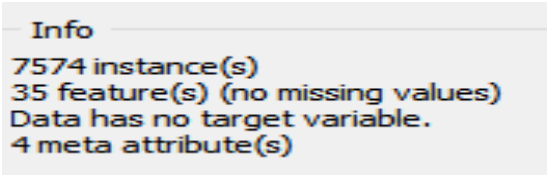
Al igual que ocurre con un modelo de regresión lineal, la regresión logística también calcula una suma ponderada de las características predictivas, pero en lugar de devolver dicho resultado, lo pasa por la función sigmoide para devolver una probabilidad. A la hora de entrenar el modelo, el objetivo es asignar altas probabilidades a las muestras positivas (aquellas para las que la variable objetivo toma el valor 1) y bajas probabilidades para las muestras negativas, lo que se consigue con la siguiente función de coste para una única instancia.

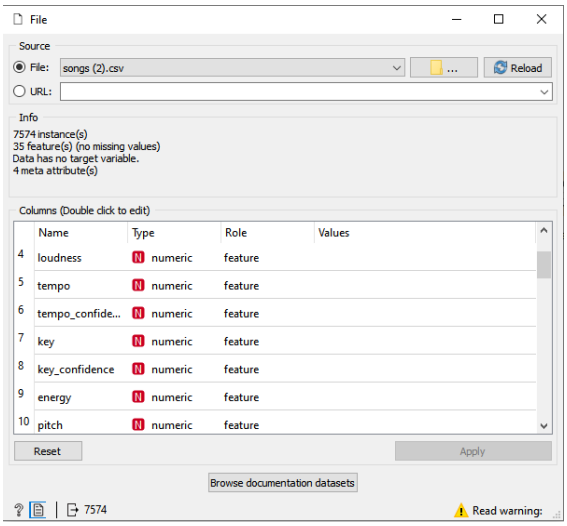
$$c(\theta) = \begin{cases} -\log(\hat{p}) & si\ y = 1 \\ -\log(1-\hat{p}) & si\ y = 0 \end{cases}$$

PREGUNTAS

Problema 1.1 - Comprensión de los datos Utilice la función File para cargar el conjunto de datos "songs.csv". ¿Cuántas observaciones (canciones) hay en total?

Para poder cargar este archivo en **orange** tenemos que tener el formato csv, así nos permitirá ver los datos que necesitamos para desarrollar el trabajo respectivo;al abrirlos encontramos





Problema 1.2 - Comprensión de los datos
¿Cuántas canciones incluye el conjunto de datos cuyo nombre de artista es "Michael Jackson"?

En esta tabla de datos se puede observar las canciones "Michael Jackson" que no se encuentra en el top 10 [figura 1] y en la [figura 2] se muestran todas la canciones que entraron en la lista del top 10 donde se encuentran las mejores canciones.

Fig 1: Canciones de Michael Jackson que no están dentro del top 10

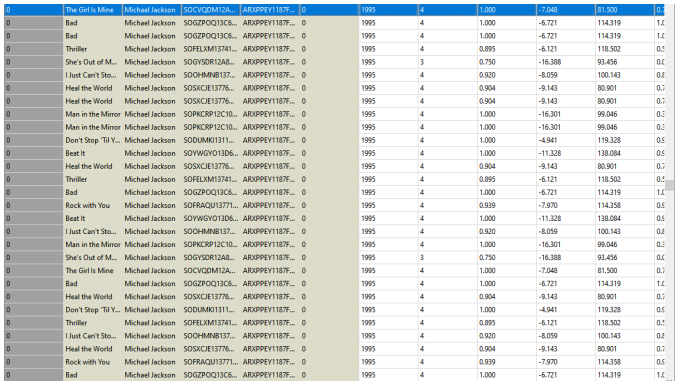
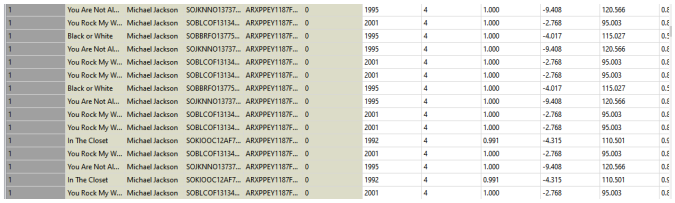
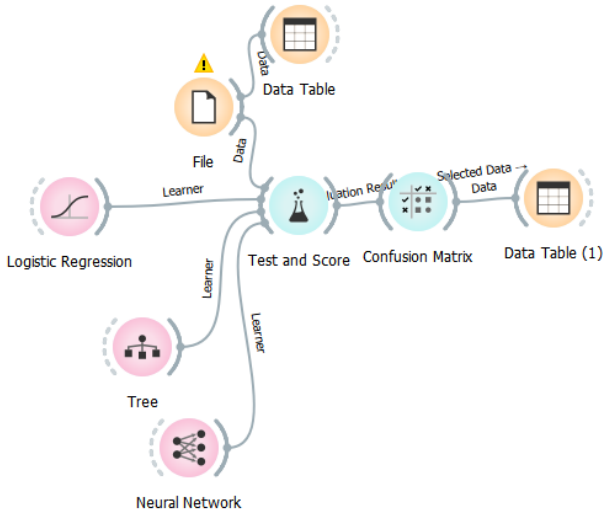


Fig 2: Canciones que están dentro del Top 10



Problema 2.1 - Creación de nuestro modelo de predicción



Regresión logística:
utilizamos el método de la regresión logística ya que fue uno de los métodos que tuvo el mayor porcentaje de acierto al momento de predecir cuáles fueron las canciones que se ubicaron en el top 10 de la lista de canciones, arrojando como resultado que el 12.8% están es este dicho top y el 87.2% fueron las canciones que no llegaron a estar en el top 10. En estas gráficas se representan los resultados obtenidos.

Regresión logística

Predicted

Actual	0	1	Σ
0	87.2 %	41.8 %	6455
1	12.8 %	58.2 %	1119
Σ	7239	335	7574

Decision Tree

Predicted

Actual	0	1	Σ
0	88.2 %	64.1 %	21950
1	11.8 %	35.9 %	3810
Σ	22526	3234	25760

Neural Network

Predicted

Actual	0	1	Σ
0	88.2 %	45.4 %	21950
1	11.8 %	54.6 %	3810
Σ	23971	1789	25760

Problema 2.3: Creación de nuestro modelo de predicción

Pensemos ahora en las variables de nuestro conjunto de datos relacionadas con la confianza del tipo de compás, la clave y el tempo (timesignature_confidence, key_confidence y tempo_confidence). Nuestro modelo parece indicar que estas variables de confianza son significativas (en lugar de las variables de firma de tiempo, clave y tempo en sí mismas). ¿Qué sugiere el modelo?

¿Cuál de estas dos opciones escogería?

- ☐ Cuanto menor sea nuestra confianza en el tipo de compás, el tono y el tempo, es más probable que la canción esté en el Top 10.
- ☑ ~~Cuanto mayor sea nuestra confianza en el tipo de compás, la clave y el tempo, es más probable que la canción esté entre las 10 mejores~~

Problema 2.4- Creación de nuestro modelo de predicción

En general, si la confianza en el compás, el tempo y la clave es baja, es más probable que la canción sea compleja. ¿Qué sugiere nuestro modelo en términos de complejidad?

¿Cuál de estas dos opciones escogería?

- [ ] Los oyentes convencionales tienden a preferir canciones más complejas.
- [x] ~~Los oyentes convencionales tienden a preferir canciones menos complejas~~

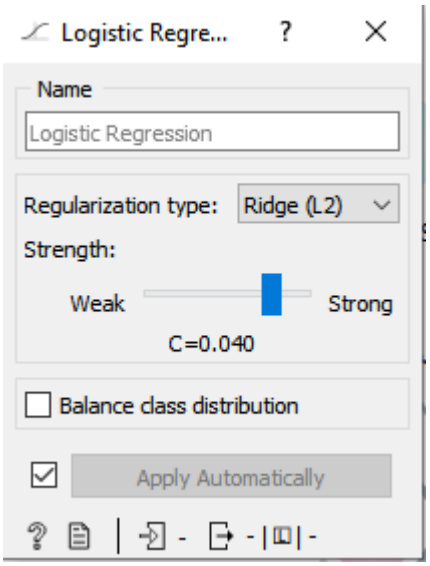
Problema 2.5 - Creación de nuestro modelo de predicción

Las canciones con instrumentación más pesada tienden a ser más fuertes (tienen valores más altos en la variable "loudness").
Al inspeccionar el coeficiente de la variable "loudness", ¿qué sugiere nuestro modelo?

¿Cuál de estas dos opciones escogería?

- [ ] Los oyentes convencionales prefieren canciones con instrumentación pesada
- [x] ~~Los oyentes convencionales prefieren canciones con instrumentación ligera~~

Problema 3.1 - Validación de nuestro modelo
Realice predicciones sobre el conjunto de pruebas utilizando nuestro modelo. ¿Cuál es la precisión de nuestro modelo en el equipo de prueba, utilizando un umbral de 0,45? (Calcule la precisión como un número entre 0 y 1.)



CONCLUSIONES

- En esta entrada se han presentado las bases de la regresión logística y su funcionamiento. Posteriormente se han descrito sus principales ventajas y desventajas. Entre sus ventajas se puede destacar su simplicidad y que sus resultados son fácilmente interpretables. Por otro lado, entre sus desventajas se puede destacar que no funciona bien en problemas que no son linealmente separables.
- La regresión logística es una técnica de aprendizaje automático relativamente sencilla. Sus resultados son interpretables y es muy usada en la industria. Funciona muy bien cuando hay muchísimos datos y las interrelaciones entre ellos no son muy complejas.
- Se espera que este documento haya contribuido a la comprensión de esta técnica, al igual que incentive la aplicación de la misma para casos del sector real por parte de los estudiantes e investigadores de temas de investigación de datos.

- Tanto los árboles de clasificación como las redes neuronales bayesianas son herramientas útiles en el análisis de la información que no se ajusta a modelos lineales, ya sea para variables que puedan ser caracterizadas como cualitativas o cuantitativas; sin embargo, con la adición de covariables (variables independientes) podría robustecerse la modelación de la información para obtener mejores resultados en etapa de prueba del modelo. En el primer caso de árboles de clasificación, el porcentaje global de clasificaciones correctas es de 0.805, mientras que en el caso de las redes regresión lógica, el coeficiente de correlación entre valores observados y predichos puede alcanzar un valor mayor de 0.826.

- Los resultados obtenidos, la cual postula a la regresión logística como una herramienta eficaz de estimación de calidad de las diferentes versiones de los proyectos. Como parte del objetivo general se ha evaluado la capacidad que tiene en la predicción de calidad de software, resultando efectivas en esta tarea y demostrando que son una herramienta útil de predicción.

BIOGRAFÍA

- https://masteres.ugr.es/moea/pages/curso201516/tfm1516/gamezortiz_tfm/!
- https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-97282016000100033
- https://www.youtube.com/watch?v=0SipBBU6sxw&t=293s
- https://www.youtube.com/watch?v=0SipBBU6sxw
- https://www.youtube.com/watch?v=lJfeDhdpcsY&t=50s
- https://www.youtube.com/watch?v=rG3J1qVA1cg&t=1118s
- https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-versatil-articulo-X0211699500035664