Regresión logística, modelos restringidos de ridge y lasso y gradiente descendiente © EDICIONES ROBLE, S.L.

Indice

Regresión logística, modelos restringidos de Ridge y Lasso y gradiente descendiente

I. Introducción

Existen modelos que además de servir para entender la relación de una variable cuantitativa continua respecto a otras, como en el caso de la regresión lineal, son capaces de estudiar la relación de una variable binaria (0-1) respecto a otras. En este caso, el objetivo es clasificar las variables dadas en una de las dos categorías.

En esta unidad se estudiará la **regresión logística**, que se encarga de explicar una variable binaria. Además, veremos cómo se pueden interpretar los coeficientes de salida y las distintas métricas que son relevantes a la hora de seleccionar un modelo de este tipo.

Finalmente, se abordarán los métodos de regularización bayesianos, que son una alternativa a los criterios estadísticos ya vistos, para seleccionar modelos.

II. Objetivos específicos

- Comprender qué es la regresión logística.
- Entender los log odds.
- Interpretar los coeficientes de la regresión logística.
- Calcular métricas de rendimiento de los modelos de clasificación 0-1.
- Aplicar modelos regularizados Lasso y Ridge.
- Entender superficialmente el algoritmo que se usa para calcular los modelos trabajados: gradiente descendente.

III. Regresión logística

Supongamos que tenemos una variable binaria que procede de una distribución de Bernouilli, de modo que puede tomar los valores 0 o 1, y toma el valor 1 con una probabilidad *p*.

Ahora supongamos que queremos conocer, dada una serie de atributos conocidos (otras variables), la probabilidad con la que esta variable binaria da positivo.

En estas circunstancias en las que la variable objetivo es una variable binaria, se usa generalmente el modelo denominado **regresión logística.** Esta es similar a la regresión lineal, pero la variable objetivo toma valores 0 o 1 y el valor de regresión que da el modelo debe ser una probabilidad *p* en el intervalo [0, 1]. La intención de este modelo es la de clasificar en una de estas dos categorías (0 o 1), aportando la probabilidad de pertenecer a cada una.

Dados la edad, el salario y el número de hijos, se requiere mostrar la probabilidad de que un cliente compre un producto concreto.

La expresión matemática de la regresión logística varía respecto a la regresión lineal:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 - \beta_2 x_2 + ... + \beta_k x_k$$

La denominación del tipo de regresión y la función que se toma en el lado izquierdo provienen de la **función sigmoide y los *logits*.** Se entiende como un *logit* la expresión que está en el lado izquierdo de la fórmula, la cual permite que el valor predicho por la regresión logística sea un número entre 0 y 1. Por tanto, un *logit* es:

$$logit(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p)$$

A su vez la función sigmoide o curva logística ocupa un lugar importante en este tipo de regresión ya que transforma cualquier número en un número en el intervalo [0,1], esto es, en una probabilidad:

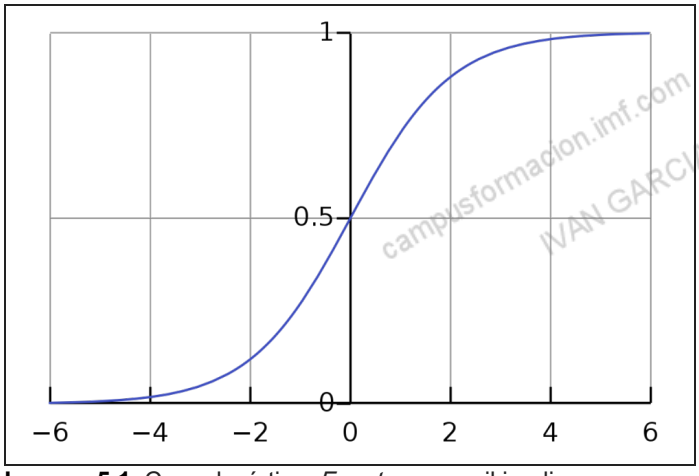

Imagen 5.1. Curva logística. *Fuente:* www.wikipedia.com

IV. Interpretación de coeficientes en regresión logística: *log odds ratio*

Suponemos que la probabilidad de éxito de un suceso es p = 0.8. Entonces el *odd ratio* es:

$$\frac{p}{1-p} = \frac{0,8}{1-0,8} = 4$$

Esto significa que hay 4 veces más probabilidad de éxito que de fracaso, ya que el *odd ratio* es el cociente entre la probabilidad de éxito (1) y la de fracaso (0).

Los logaritmos y la expresión definida de la regresión logística se transforman en los *log odds ratios*, esto nos ayuda a interpretar los coeficientes de la regresión logística.

Que un factor tenga por coeficiente β significa que por cada unidad de aumento de ese factor se verifica que la probabilidad de éxito se multiplique por e^{β}

Ejemplo: en el caso dado en la descripción de la regresión logística, se supone que el coeficiente del número de hijos es 0.1. Entonces, el aumento de probabilidad de comprar el producto por cada hijo es de:

$$e^{\beta} = e^{0,1} = 1,105171$$

Esto es, por cada hijo la probabilidad de que compre el producto aumenta en un 10.51%.

De este modo, se puede interpretar la influencia de los factores en la probabilidad final de que la variable binaria sea un 1.

V. Métricas en problemas de clasificación binaria

Los problemas de clasificación 0-1 son los más abundantes en el entorno de negocio. Aplicar modelos estadísticos como la regresión logística tiene dos objetivos:

- Explicar la variable objetivo respecto a los predictores, aportando un mayor entendimiento de los datos que se plantean.
- Realizar predicciones.

Las predicciones no son perfectas y tienen errores. Además, en función de la naturaleza de la variable objetivo interesará optimizar una métrica u otra.

Cuando generamos un clasificador binario, este tiene asociada una matriz de confusión que indica los valores de acierto y error en cada clase real:

	Predicciones	
	Positivo	**Negativo**
Positivo en la realidad	TP	FN
Negativo en la realidad	FP	TN

Tabla 5.1. Valores de acierto y error. *Fuente:* elaboración propia.

Las descripciones de las componentes de la matriz de confusión son:

TP

Cantidad de predicciones positivas que son realmente positivas.

FP

Cantidad de predicciones positivas que son realmente negativas.

TN

Cantidad de predicciones negativas que son realmente negativas.

FN

Cantidad de predicciones negativas que son realmente positivas.

Métricas:

Accuracy

Precisión global a través de todas las clases.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precission

Capacidad de acierto dentro de la clase predicha como positiva.

$$\frac{TP}{TP + FP}$$

Recall

Capacidad de acierto dentro de la clase positiva.

$$\frac{TP}{TP + FN}$$

F1-score

Media armónica entre precission y recall. Es una métrica equilibrada entre tener pocos falsos positivos y pocos falsos negativos.

$$2\frac{Precision * Recall}{Precision + Recall}$$

Existen más métricas y se pueden adaptar a cada caso de aplicación. Se han mostrado las fundamentales para entender cómo se mide la efectividad de un clasificador.

Ejemplos de usos:

- Diagnosticar un tumor como maligno o benigno en función de una colección de lecturas médicas. En este caso, el objetivo es no mandar pacientes enfermos a casa, por tanto, se maximizará el **recall**.
- Predecir si una transacción es fraudulenta o no. En este caso, el objetivo es reconocer fraude, pero sin molestar a los clientes que no han cometido ninguna ilegalidad. Se maximizará, entonces, la **precission**.
- Predecir si una reserva de un hotel se cancelará o no. En este caso, se persigue un equilibrio entre reconocer cancelaciones y aplicar promociones, y aplicar promociones a reservas que no se van a cancelar, por tanto, el **f1-score** será la métrica a maximizar.

Descarga: Consulta el notebook UD5 N01

Descárgate el archivo UD5_N01 y ejecútalo en R. También puedes verlo en.html

VI. Modelos restringidos de Ridge y Lasso

Para seleccionar el modelo y las variables que participan en él, se han visto criterios como AIC y BIC, basados en verosimilitud con penalización por complejidad.

Otra vía para establecer el mejor modelo de representación de los datos son los **modelos regularizados**. Estos están basados en conceptos de **probabilidad bayesiana** y en la **navaja de Ockham**: "ante rendimientos comparables, el modelo que explica la realidad de manera más simple es preferible".

Los modelos de Ridge y Lasso atienden a ello, introduciendo unas penalizaciones en los valores de cada coeficiente, que hacen que el valor natural sea 0.

Al contrario que en AIC o BIC —los cuales son utilizados si una variable entra o no entra de manera discreta (SÍ o NO)—, en estos modelos se produce una restricción continua sobre los valores de los coeficientes de cada variable que hace que estos sean más interesantes y potentes como modelos de predicción.

Además, los modelos regularizados de Ridge y Lasso afrontan el problema de la **multicolinealidad**. Esta se presenta cuando existen variables predictoras que son casi dependientes entre sí, lo que produce un efecto de inestabilidad del resultado numérico del modelo, si no se trata adecuadamente.

Los modelos tienen las siguientes cualidades:

Modelo con regularización Ridge

El valor natural de los coeficientes es 0, penalizando la adjudicación de valor. En este modelo no se obtienen coeficientes finales nulos, aunque sean muy pequeños. Es, por tanto, un modelo que regulariza de manera continua todos los coeficientes.

Modelo con regularización Lasso

El valor natural de los coeficientes es 0, penalizando la adjudicación de valor en cada uno. Sin embargo, a diferencia de Ridge y por la construcción de esta regularización, este modelo establece coeficientes finales con valor nulo (los que no son importantes). Los modelos de tipo Lasso se conocen como **modelos huecos o *sparse*** debido a la selección de variables que se produce al establecer coeficientes como nulos.

Ambos modelos se construyen dependiendo de una **constante de regularización**, usualmente nombrada como λ. Esta constante toma valores positivos. Si se adopta λ=0, el modelo no estará regularizado y equivaldrá a una regresión lineal o logística simple.

Conforme λ es mayor, el modelo estará más regularizado, esto es: los coeficientes son más próximos a 0 en el caso Ridge y hay más coeficientes nulos en el caso Lasso.

Es importante destacar que los modelos restringidos se aplican tanto en el caso de regresión lineal como en el de regresión logística vistos.

Descarga: Consulta el notebook UD5 N02

Descárgate el archivo UD5_N02 y ejecútalo en R. También puedes verlo en .html

VII. Algoritmo de gradiente descendente

El ajuste de modelos estadísticos generalmente se establece mediante la **maximización** de una **función de verosimilitud** asociada. Esta función representa de alguna forma la probabilidad de que se sucedan los datos dispuestos bajo las suposiciones del modelo.

Por tanto, cuando se realiza este ajuste, se buscan los valores de los parámetros del modelo que hacen que el valor de esta función sea mayor. En el caso de modelos como los de Ridge y Lasso, se incluyen términos aditivos de penalización de complejidad que actúan como regularizadores.

$$\hat{\theta} = argmax F_{verosimilitud}(\theta)$$

Donde θ es el conjunto de parámetros del modelo.

Esta maximización se plantea como una ecuación, pero no se resuelve como una operación algebraica directa, ya que por problemas de estabilidad numérica es, en general, imposible o indeseable porque da resultados imprecisos.

En general, se utilizan **algoritmos iterativos** que generan una secuencia de valores de los parámetros que converge (se acerca) al valor óptimo si se verifican una serie de condiciones.

De entre toda la familia de algoritmos iterativos útiles para hallar los valores óptimos, los **algoritmos iterativos voraces** son los preferidos por su sencillez de planteamiento, velocidad y efectividad. Los algoritmos voraces son algoritmos que en cada iteración dan un paso en la dirección que localmente queda establecida como mejor para conseguir el objetivo.

El **gradiente descendente** es el mejor ejemplo de algoritmo iterativo voraz. Plantea el espacio de valores de los parámetros como un mapa y el valor de la función de verosimilitud como la altura. Busca, en cada paso, la dirección en la que el valor de la función de verosimilitud aumenta más (dirección de máxima pendiente de una montaña) y se mueve por el espacio de parámetros de esta forma.

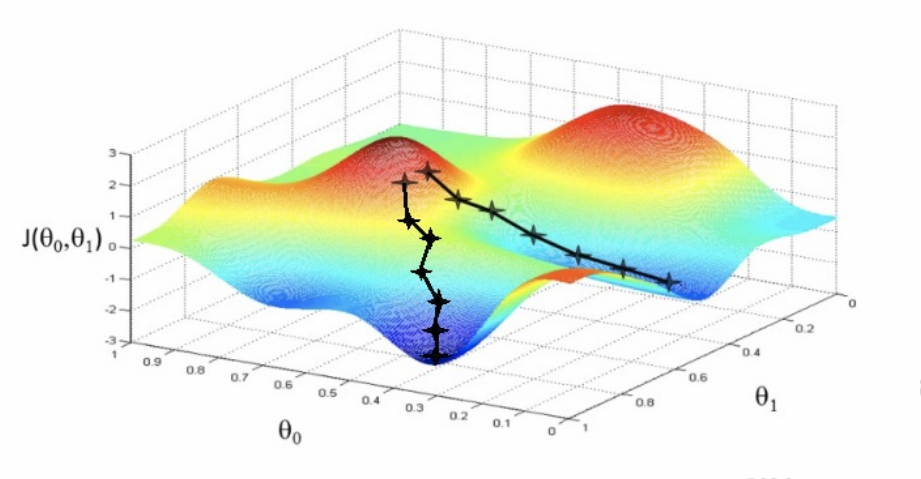

Imagen 5.2. Algoritmo de gradiente descendente.
Fuente: https://www.youtube.com/watch?v=5u4G23_OohI

Si estamos en una zona montañosa y queremos subir a la cima más alta, un algoritmo voraz sería que cada paso se diera en la dirección hacia la cual la pendiente es máxima. Esto es el resumen intuitivo del algoritmo de gradiente descendente (algoritmo iterativo voraz).

VIII. Resumen final

Se ha visto el modelo de clasificación fundamental: la regresión logística. También se ha mostrado cómo se interpretan sus coeficientes y las métricas fundamentales para evaluar su rendimiento.

Por otra parte, se ha estudiado el algoritmo central del aprendizaje estadístico, ya que es eje de acción para hallar los coeficientes en muchos de los modelos.

Además, se han visto los modelos regularizados de Ridge y Lasso, aplicables tanto a regresión lineal como logística. Estos modelos dan resultados muy sólidos de predicción a la vez que son interpretables, por lo que suponen una herramienta de trabajo más que suficiente para interpretar los datos y aplicar en predictivo.

Ejercicios

Caso práctico

Como repaso del tema y preparación para el Caso práctico final, se presenta el siguiente caso práctico.

Descárgate el archivo ACTIVIDAD5_UD5 en R y la csv del caso. También puedes verlo en .html

Cuando lo hayas realizado, puedes descargar su solución y comprobar tus resultados.

Solución

En los siguientes archivos dispones de la solución de la actividad propuesta:

- Solución en .html
- Solución en R.

Recursos

Glosario.

- **Algoritmo iterativo voraz**: Algoritmo que genera una secuencia en la que en cada paso se toma la decisión que localmente sea mejor. Se espera que converja a un óptimo local al menos.

- **Log odd**: Cada una de las probabilidades asociadas a cada variable predictora de una regresión logística, que indica cómo se multiplica la probabilidad de valer 1 en la variable objetivo por una unidad de aumento de la variable predictora.

- **Métrica**: Medida evaluadora del rendimiento de un modelo.

- **Modelo de clasificación 0-1**: Estructura estadística que sirve para relacionar una variable binaria con otras variables predictoras.

- **Modelo regularizado**: Es un modelo que incluye una penalización en la atribución de coeficientes intrínseca. De este modo, cuando el algoritmo de gradiente descendente calcula los coeficientes, ya tiene en cuenta el equilibrio sesgo-varianza.

- **Variable binaria**: Variable que solo toma los valores 0 y 1.