Práctica 5

1. Estrategias de validación

Cualquier modelo estadístico es generado por un conjunto de observaciones (conjunto de entrenamiento) y su finalidad es predecir una (o varias) variable de respuesta para observaciones nuevas (conjunto de testeo). El error en la predicción es lo que se usa habitualmente para evaluar que “tan bueno” es el modelo. Para que el modelo se ajuste mejor a los “datos nuevos”, se utilizan las estrategias de validación.

Validación simple

Es la división de los datos en dos conjuntos, de forma aleatoria, un grupo para entrenamiento y otro para testeo.

Leave One Out Cross Validation (LOOCV)

Este método consiste en tomar todos los datos como de entrenamiento, menos uno, que se utiliza para evaluación. Cómo el error dependerá de la observación seleccionada para evaluar, este proceso se repite tantas veces como observaciones el conjunto de datos, excluyendo cada vez una observación distinta.

El error estimado en LOOCV es el promedio de todos los errores calculados en cada iteración.

K-Fold Cross Validation

En k-fold CV, también se itera, pero en un número definido de veces: las k particiones. De ellas, k-1 se utilizan para entrenar el modelo y la restante para evaluarlo. En este proceso se encuentran k estimaciones del error, que se promedian para obtener el error final.

Habitualmente se elige k = 5 o 10. Dependerá de la cantidad de observaciones que se disponga en los datos.

Implementación en R

Para conseguir una partición de determinado tamaño de un conjunto de datos dado, vamos a utilizar la función *sample()* del paquete base de R. También se puede utilizar la función *sample.split()* del paquete **caTools** [2].

Las funciones utilizan parámetros distintos para definir la partición:

- En el caso de *sample()*, se especifica **size** como un número entero, que indica la cantidad de elementos a elegir.

- Para *sample.split()* se especifica el porcentaje de los datos con **SplitRatio**.

Para el caso de LOOCV, se puede hacer la validación de forma manual, incluyendo el modelo a validar dentro de un ciclo o utilizar funciones definidas en R para tal fin. También se puede utilizar la opción “LOOCV” como método de validación en los parámetros de control del entrenamiento de modelos para el paquete **caret** [3], utilizando las funciones *trainControl()* y *train()*.

Por último, k-fold CV se puede realizar utilizando las mismas funciones del paquete **caret**, pero con las opciones “cv” y definiendo el valor de k; y también se puede implementar de forma manual.

Muchos modelos implementados en R tienen su función para k-fold CV.

2. Regresión Logística

La regresión logística, es un método de regresión para estimar la probabilidad de una variable binaria o dicotómica, que generalmente codificaremos con 1 (la característica de interés) y 0. Entonces, vamos a estar resolviendo un problema de clasificación binaria.

Cuando planteamos regresión lineal, como la relación más simple entre dos variables, teníamos:

$$y = m \times x + b$$

Este modelo tiene varias cuestiones por las que no se puede aplicar para intentar estimar una respuesta binaria. La primera, ***y*** no tiene distribución normal; segundo, no acota $p(x)$ y $0 < p(x) < 1$. Entonces, solamente con estas dos consideraciones, no podríamos utilizarla.

Por estos problemas del modelo lineal, se busca un modelo alternativo de la forma:

$$Y = F\left(\alpha + \beta x\right) + \epsilon(x)$$

Siendo $\epsilon(x)$ variables aleatorias independientes con esperanza 0, por lo que podemos escribir:

$$p(x) = F(\alpha + \beta x)$$

Con ***F*** una función monótona creciente. La ecuación anterior se puede expresar como:

$$F^{-1}\left(p(x)\right) = \alpha + \beta x$$

Entonces, buscamos una función ***F***, cuya inversa aplicada a *p(**x**)* modele linealmente la relación con la variable predictora. La más utilizada es la transformación *logit*:

$$logit(p(x)) = \ln \frac{p(x)}{1 - p(x)}$$

Que expresándola en función de *p(**x**)* nos queda:

$$p(x) = \frac{\exp\left(\alpha + \beta x\right)}{1 + \exp\left(\alpha + \beta x\right)}$$

Al igual que la regresión lineal, la regresión logística puede ser simple o múltiple. Dependiendo de la cantidad de variables predictoras que utilicemos.

Implementación con R

Para generar un modelo de regresión logística, vamos a utilizar la función *glm()* del paquete **stats** de la instalación base de R, ya que no es más que una generalización de los modelos lineales. Los parámetros que vamos a utilizar son:

- formula: vamos a indicar que variable va a ser la dependiente y cuáles las explicativas, con el operador **~**.
- data: el data.frame del que se especifican las variables (opcional).
- family: una cadena con el nombre de la función que vamos a utilizar para nuestro modelo. En este caso, el valor es **binomial**.

Para predecir utilizando el modelo encontrado con *glm()*, se utiliza la función *predict()*. Aunque se debe agregar el parámetro **type** con el valor “response” para que devuelva las probabilidades predichas.

Ejercicios

Ejercicio 1 – Un estudio quiere establecer un modelo que permita calcular la probabilidad de obtener una beca para la universidad, a alumnos en el último año de la escuela secundaria, en función de la nota que obtuvieron en matemática. La variable beca está codificada como 0 si no aplica a la beca y 1 si aplica. Los datos del estudio se encuentran disponibles en el archivo "beca.csv".

Utilice validación simple del tipo 70%-30%, 80%-20% y 90%-10% para encontrar el modelo. Grafique y evalúe los modelos con los datos de test mediante la curva ROC y el AUC. ¿Con que porcentaje de validación obtiene el mejor resultado?

Ejercicio 2 – Cargue el dataset "Default", del paquete ISLR de R [4], que contiene datos de la falta de pago de las deudas de la tarjeta de crédito de 10.000 clientes. Las variables son:

- default: indica si el cliente es moroso o no.
- student: indica si el cliente es estudiante o no.
- balance: saldo promedio de la tarjeta de crédito después de realizar el pago mensual.
- income: ingresos del cliente.

A partir de estos datos, resuelva utilizando validación simple:

a. Explore los datos. ¿De qué clase son las variables relevadas?

b. Encuentre un modelo de regresión logística simple, para predecir la probabilidad de falta de pago de los clientes, utilizando como predictor los ingresos del cliente.

c. Grafique los datos y el modelo encontrado. Evalúe el modelo.

d. Ahora utilice el saldo de la tarjeta de crédito como variable predictora. ¿Esta variable le permite mejorar la estimación de los clientes morosos?

Ejercicio 3 – Según datos relevados de la UNER (archivo "c_consulta.csv"), se supone que existe una relación entre el hecho de que un estudiante asista a clases de consulta de estadística (Si = 1, No = 0), la nota obtenida en el primer parcial de la materia y el sexo del estudiante.

a. Explore los datos. De qué tipo son las variables?

b. Genere un modelo con las variables adecuadas que prediga la probabilidad de que el estudiante tenga que asistir a clases de consulta.

Evalúe el modelo encontrado y encuentre el porcentaje de datos bien clasificados.

Referencias

1. “Estadística y Machine Learning con R”. Disponible en: https://bookdown.org/content/2274/portada.html
2. “Package caTools”. Disponible en: https://cran.r-project.org/web/packages/caTools/caTools.pdf
3. “Package caret”. Disponible en: https://cran.r-project.org/web/packages/caret/caret.pdf
4. “Package ISLR”. Disponible en: https://cran.r-project.org/web/packages/ISLR/ISLR.pdf
5. “Package pROC”. Disponible en: https://cran.r-project.org/web/packages/pROC/pROC.pdf