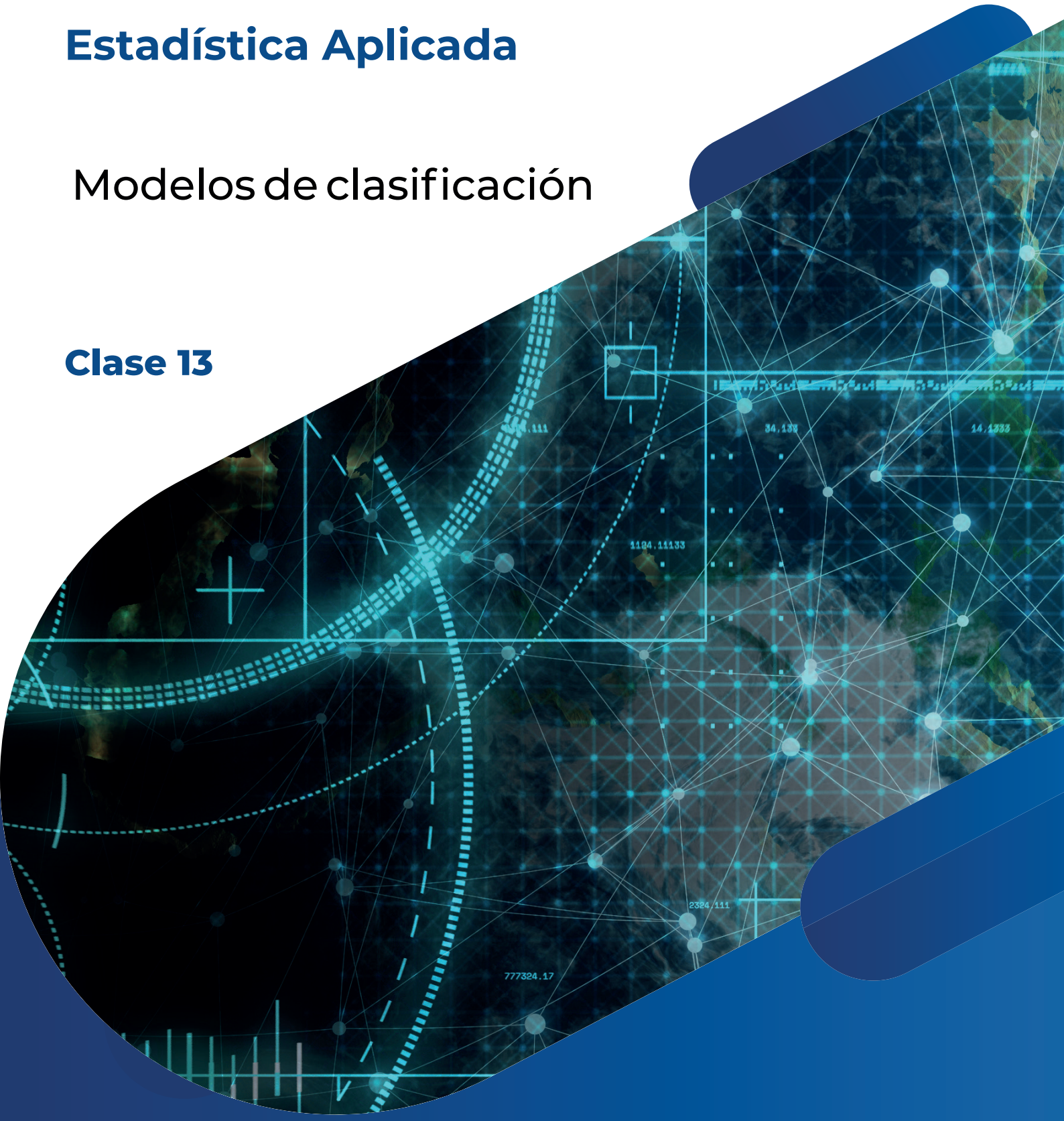
Estadística Aplicada

Modelos de clasificación

Clase 13



Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, trataremos sobre modelos de clasificación. Hasta el momento, todos los modelos fueron desarrollados con variables de respuesta continuas; sin embargo, los modelos de clasificación son más comunes. Existe una gran cantidad de modelos de clasificación en la ciencia de datos, pero en esta clase, nos centraremos específicamente en modelos de regresión logística. Adicionalmente, en este punto introduciremos algunas pruebas de asociación utilizadas cuando las variables categóricas no son dicotómicas.

Por otro lado, en estos modelos, continuamente se usan variables categóricas, por lo que es importante entender cómo este tipo de variables pueden ser codificadas de tal manera que se pueda obtener un mejor rendimiento del modelo, al permitir codificar variables categóricas como variables numéricas o cuasi numéricas.

Clase 13:

El resultado o resultados de aprendizaje que serán abordados con el contenido de la clase son:

Analizar información contextual sobre hábitat, infraestructura y movilidad, mediante técnicas estadísticas descriptivas e inferenciales, para la adecuada toma de decisiones

Reto # 4

13) Modelos de clasificación

Cuando hablamos de modelos cuya variable de respuesta es discreta o categórica, nos referimos a modelos de clasificación. Si la variable de respuesta solo puede tomar dos valores (verdadero/falso o 0/1), nos referimos a ellos como modelos binomiales; por el contrario, si puede tomar un conjunto de valores, nos referimos a ellos como modelos multinomiales. En la presente clase veremos modelos logísticos binomiales.

13.1) Modelos logístico

La regresión logística es un modelo que se utiliza para determinar la probabilidad de que ocurra un evento. Estos modelos muestran la relación entre variables y luego calculan la probabilidad de un resultado determinado. Para la creación de este tipo de modelos, recurrimos a la función glm().

Hasta ahora, hemos utilizado la función lm() para generar modelos que tienen determinados supuestos, por ejemplo, la distribución debe ser gaussiana o normal, y su varianza debe ser homocedástica. Pero no todos los problemas cumplen con estos supuestos. ¿Qué pasa si los datos reflejan una distri-

bución binomial o de Poisson? Para esto, introduciremos la función glm.

Modelos lineales generalizados

Explica el uso de glm

glm().

Los modelos GLM son una generalización de los modelos lineales vistos, que permiten crear modelos en los cuales algunas condiciones puedan ser especificadas. Una de estas condiciones, supone que cada resultado Y de las variables de respuesta se genera a partir de una distribución particular en la familia exponencial, esta incluye la distribución normal, binomial, Poisson y gamma, entre otras.

Los modelos glm tiene 3 componentes:

Componente aleatorio: : Especifica la distribución de probabilidad de la variable de respuesta. Específicamente, para una distribución dada (normal, binomial, etc.), estiman su media y varianza. No hay un término de error separado (ε) como en la regresión lineal.

Componente Sistemático: Especifica las variables explicativas del modelo y su combinación lineal. Esto es similar a la regresión lineal.

Función de enlace: Indica cómo el valor esperado de la respuesta se relaciona con la combinación lineal de variables predictoras. Glm NO asume una relación lineal entre la variable de respuesta y las variables predictoras (como lm), pero sí asume una relación lineal entre la respuesta esperada transformada en términos de la función de enlace y las variables explicativas. Específicamente, para regresión logística utilizaremos la familia binomial (family=binomial(link = “logit”))

Note que la distribución está dada por distintas “familias” y entre paréntesis aparece la función “link”, que indica justamente la función de enlace que se usará en el proceso.

La familia de binomial dispone de varias funciones de enlace que permiten el comportamiento requerido. La siguiente tabla presenta el nombre de la función y su fórmula de cálculo.

Tabla 1

Funciones de enlace para regresión logística

Creación del autor Alfonso Prado

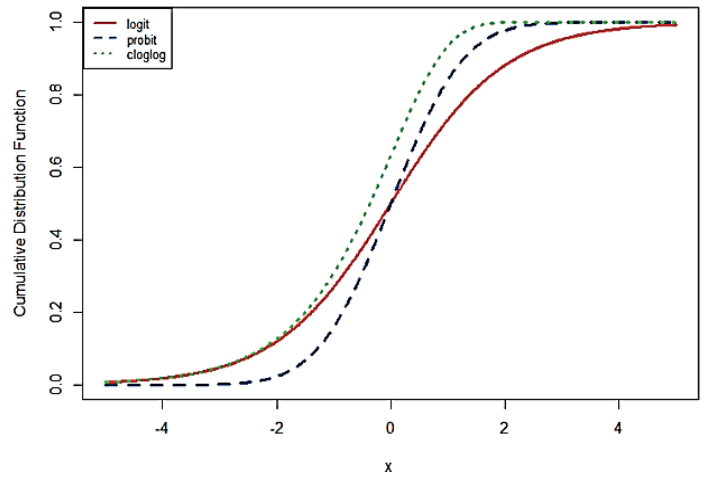
Nombre	Función	Probabilidad
Logit	$z = \log\left(\frac{p}{1-p}\right)$	$p = \frac{e^z}{1+e^z}$
Probit	$z = \Phi^{-1}(p)$	$p = \Phi(z)$
Cauchit	Na	$p = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$
Cloglog	Na	$p = 1 - e^{-e^z}$

Las funciones de enlace tienen distintas características, como se evidencia en la siguiente figura.

Figura 1

Características de funciones de enlace

Creación del autor Alfonso Prado



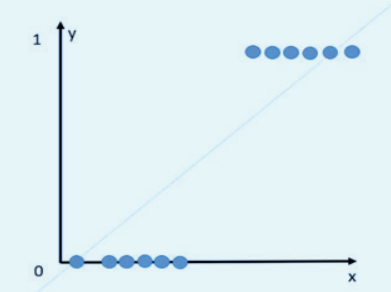
Por defecto, en el caso de distribución binomial, la función es “logit”, que, como vemos en la tabla 1, se basa en el concepto de Odds.

La siguiente figura explica el concepto de la función de enlace. Para un cierto rango de valores de la variable predictora X, se desea que la respuesta Y sea 0. Sin embargo, cuando X sube de cierto valor, se espera que Y sea 1.

Figura 2

Relación entre X y Y en un modelo logístico

Creación del autor Alfonso Prado



13.2) Medición de odds y odds-ratio

El análisis de datos categóricos se basa típicamente en tablas de contingencia de dos o más dimensiones, tabulando la frecuencia de ocurrencia de niveles de datos nominales y/o ordinales. Una tabla de contingencia es una herramienta utilizada para crear al menos dos filas y dos columnas que representan datos categóricos en términos de conteos de frecuencia. La tabla permite medir la interacción entre dos variables para conocer una serie de información "oculta" de gran utilidad y comprender con mayor claridad los resultados de una investigación.

Por ejemplo, tenemos un dataset que presenta la ocurrencia de una enfermedad (variable mf), desglosada por otras variables como área, rango de edad, sexo, y queremos entender cómo se distribuye la variable mf de acuerdo con el área (o cualquier otra variable categórica).

Figura 3

Dataset (a) y tabla de contingencia (b)

Creación del autor Alfonso Prado

(a)

id ID	mf Infection	area Residence	agegrp Age group (years)	sex Sex
1	Infected	Savannah	20-39	Female
2	Infected	Rainforest	40+	Male
3	Infected	Savannah	40+	Female

(b)

	Residence		
	Savannah, N = 548	Rainforest, N = 754	Overall, N = 1,302
Infection			
Not-infected	267	213	480
Infected	281	541	822

En principio, la tabla de contingencia toma el nombre de las filas y columnas de las variables que anal-

iza. Sin embargo, se acostumbra a crear la tabla con nombres genéricos como Exposure y Outcome, donde las filas (Exp+ y Exp-) indican haber estado sujeto a un "tratamiento" y las columnas (Out+, Out-) haber desarrollado un efecto o no. Debemos tomar en cuenta que la cantidad de observaciones de un valor nominal no va a ser necesariamente la misma de otro valor. Por ejemplo, la cantidad de encuestados hombres (Male) no es igual a las mujeres (Female), por lo tanto, la tabla debe ser procesada a fin de encontrar el valor del odds. El Odds nos indica la probabilidad entre Exposure y Outcome. Ejemplo: En la tabla siguiente odds de los expuestos=140/84, odds de no expuestos=139/92.

Tabla 2

Odds y Odds Ratio

Creación del autor Alfonso Prado

	Outcome +	Outcome -	Total	Inc risk *	Odds
Exposed +	140	84	224	62.5	1.67
Exposed -	139	92	231	60.2	1.51
Total	279	176	455	61.3	1.59

Point estimates and 95% CIs:

Inc risk ratio	1.04 (0.90, 1.20)
Odds ratio	1.10 (0.76, 1.61)

Sin embargo, los odds de cada grupo no son suficientes para obtener una idea de si la variable es significativa. Esto se debe a que, del total de expuestos, no todos desarrollaron un resultado positivo. Lo mismo se puede decir a la inversa de los no expuestos. Por lo tanto, para comparar los dos grupos (expuestos y no expuestos), dividiremos los valores de cada grupo y esto se denomina el odds ratio o OR. Del ejemplo, esto sería igual a 1.67/1.51 =1.1

¿Cómo interpretar el odds-ratio? La razón de probabilidades nos dice cuánto más altas son las probabilidades de exposición entre los casos de una salida positiva que entre los casos de una salida negativa. Mientras mayor sea la razón de los odds (OR), mayor será la probabilidad de que los eventos se asocien positivamente. Si la razón de probabilidades es menor que 1, los eventos se asocian negativamente. Si es cercana a 1, no hay influencia. Podemos entender esto como una correlación entre variables nominales. Note que se divide la probabilidad de que se presente el evento para la probabilidad de que no se presente, cuyo logaritmo es utilizado por la función logit (Ver tabla 1)

Figura 4

Odds ratio como relación de probabilidad

Creación del autor Alfonso Prado

$$\pi_1/(1-\pi_1)$$

Veamos un ejemplo:

Se trata de un dataset de una empresa que emplea, en un momento dado, a unos 4000 empleados. Sin embargo, cada año, alrededor del 15% de sus empleados abandona la empresa y necesitan ser reemplazados. Esto ocasiona una serie de problemas a la empresa, por lo que se recurre a un modelo logístico que ayuda a predecir la deserción de los empleados. Dentro de las variables predictoras, tenemos el género y la respuesta es attrition. ¿Cuál es la relación entre género y attrition?

#Obtenemos la tabla de contingencia

table(data$Gender, data$Attrition)

```
            No  Yes
Female    1464  265
Male      2141  430
```

Procesamos la tabla de contingencia

epi.2by2(table(data$Gender, data$Attrition))

```
             Outcome +    Outcome -      Total      Inc risk *        Odds
Exposed +         1464          265       1729            84.7        5.52
Exposed -         2141          430       2571            83.3        4.98
Total             3605          695       4300            83.8        5.19

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 1.02 (0.99, 1.04)
Odds ratio                                     1.11 (0.94, 1.31)
Attrib risk in the exposed *                   1.40 (-0.83, 3.63)
Attrib fraction in the exposed (%)            1.65 (-0.99, 4.22)
Attrib risk in the population *                0.56 (-1.25, 2.38)
Attrib fraction in the population (%)         0.67 (-0.40, 1.73)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 1.492 Pr>chi2 = 0.222
Fisher exact test that OR = 1: Pr>chi2 = 0.237
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

El análisis nos indica que los empleados varones tienen un 11% más de probabilidades que renunciar.

13.3 Prueba Chi-cuadrado de independencia

El proceso de tablas de contingencia visto anteriormente funciona bien para variables dicotómicas.

Pero ¿qué pasa si tenemos más de dos niveles de la variable? Por ejemplo, del mismo dataset mencionado anteriormente, existen variables como Marital Status (estado civil), que tiene tres niveles (casado(a), divorciado(a), soltero(a)). En este caso, la función 2by2 no funciona, por lo que debemos recurrir a otra prueba de independencia conocida como Chi-cuadrado o Ji-cuadrado.

Tabla 3

Tabla de contingencia MaritalStatus – Attrition

Creación del autor Alfonso Prado

	No	Yes
Divorced	855	94
Married	1718	251
Single	1032	350

En estos casos, necesitamos realizar los siguientes pasos:

a) Mostrar que existe una relación para lo cual usaremos la prueba de independencia Xi^2 (se pronuncia chi-cuadrado).
b) La prueba Xi^2 (al igual que otras pruebas) consiste en 2 partes, obtener un estadístico y compararlo con su distribución. En este caso, la distribución es Chi cuadrado, y los grados de libertad se calcula como (filas -1)*(columnas -1) de la tabla de contingencia
c) Validar en forma gráfica mediante diagramas de mosaicos
d) Digitalizar o codificar la variable

De acuerdo con Ramírez-Alan ,(2016) "El test Xi^2 considera la hipótesis nula (H0) de que las variables son independientes. Si esto es verdad, la frecuencia de ocurrencia debería estar dada por la cantidad de casos totales multiplicada por la probabilidad esperada." En el ejemplo anterior, si la probabilidad de renunciar es igual para los 3 niveles (π_{ij} = 33%) entonces, si se han detectado 1000 renuncias, estas deberían estar distribuidas en las 3 categorías ($n*\pi_{ij}$ =1000*.33). Este valor conocido como μ_{ij}.

Figura 5

Frecuencia esperada

Creación del autor Alfonso Prado

Frecuencia esperada $\mu_{ij} = n\ \pi_{ij}$

Pero si Ha está en lo correcto va a existir una diferencia entre frecuencia observada (ηij) y la espera-

da (μij lo que indicaría que existe algún fenómeno subyacente que influye en la frecuencia observada. La prueba χ^2 (chisq.test) calcula esta diferencia de la siguiente forma:

Figura 6

Fórmula de Xi^2

Creación del autor Alfonso Prado

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

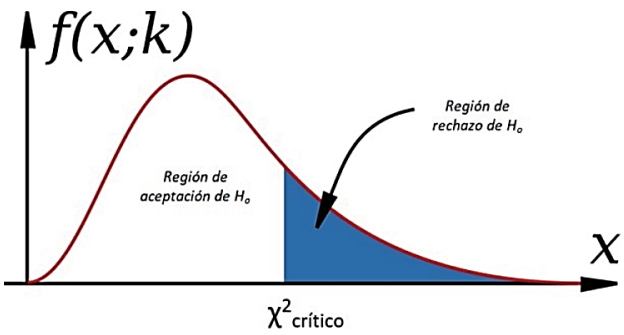
Mientras mayor sea este indicador, mayor las probabilidades de que las variables estén relacionadas, por el contrario, el numerador tiende a 0 .

La prueba retorna un p-value que debe ser contrastado contra un Xi-crítico, obtenido en base a la función qchisq. La siguiente figura explica las zonas de aceptación y rechazo de la asociación entre variables.

Figura 7

Zonas de aceptación y rechazo de la prueba Xi^2

Creación del autor Alfonso Prado



Continuando con el ejemplo

```
#Obtenemos la tabla de contingencia de MaritalStatus

tab <- table(data$MaritalStatus, table(data$Attrition)



             No  Yes
 Divorced   855   94
 Married   1718  251
 Single    1032  350


#Ejecutamos la prueba Chi²

chisq.test(tab)



        Pearson's Chi-squared test

data:  observ
X-squared = 130, df = 2, p-value < 2.2e-16



#calculando chi-crítico

chi_critico <- qchisq(.05, df=2, lower.tail=FALSE)

chi_critico

[1] 5.991465
```

Al obtener un estadístico chi² de 130 con un chi-critico de 5.99, concluimos que existen diferencias sustanciales, lo que indica que la variable es significativa.

Validación Gráfica

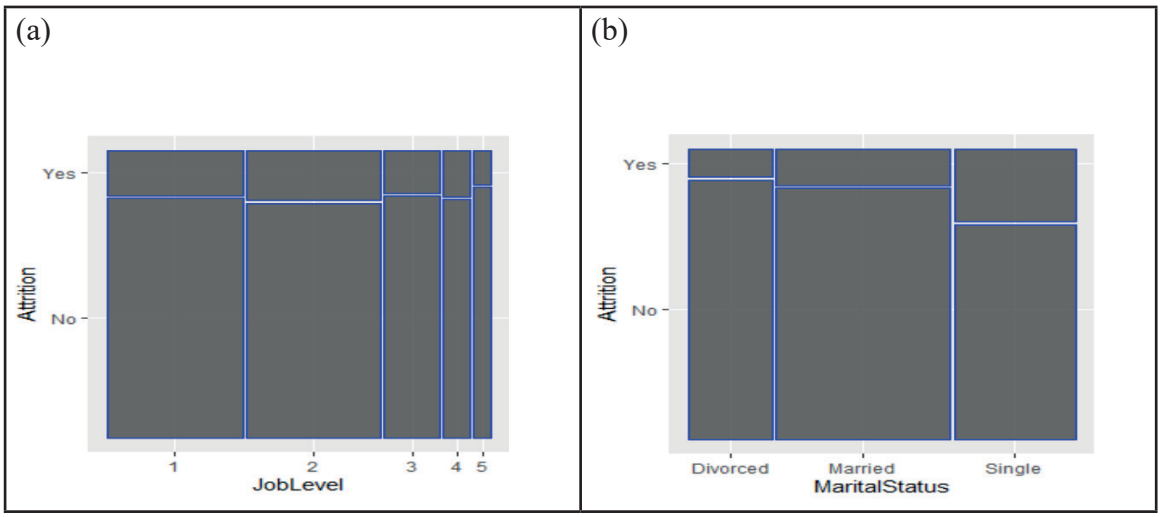
Una forma alternativa de validar la significancia de las variables categóricas (factores) con múltiples niveles es mediante un diagrama de mosaico. En estos diagramas, el área es directamente proporcional a la frecuencia de ocurrencia en la tabla, y debemos interpretarla de la siguiente manera: si obtenemos un diagrama muy "regular", se indicaría que las dos variables no están relacionadas. La palabra "regular" debemos entenderla como que las proporciones se mantienen constantes. Por otro lado, si el diagrama es irregular, indicaría que algún factor subyacente está afectando la relación con la variable predictora y, por lo tanto, esta variable puede ser significativa para un modelo logístico.

Los siguientes diagramas de mosaico visualizan la relación entre dos variables. La figura 7a muestra independencia cuando el diagrama es regular; en cambio, figura 7b) muestra un diagrama irregular, que demuestra que hay una relación.

Figura 8

Diagramas de mosaico para mostrar relacionamiento, variable no significativa (a) y variable significativa (b)

Creación del autor: Alfonso Prado



Codificación de variables categóricas (parte I)

Un modelo logístico puede incluir tanto variables continuas como categóricas o nominales. De hecho, tanto glm como lm en R implementan por defecto algún tipo de codificación para variables nominales, pero en otros lenguajes, el analista debe preparar los datos con la codificación adecuada para su uso.

Codificación de variables

Explica el uso en modelos

Enlace

Codificación por etiquetas

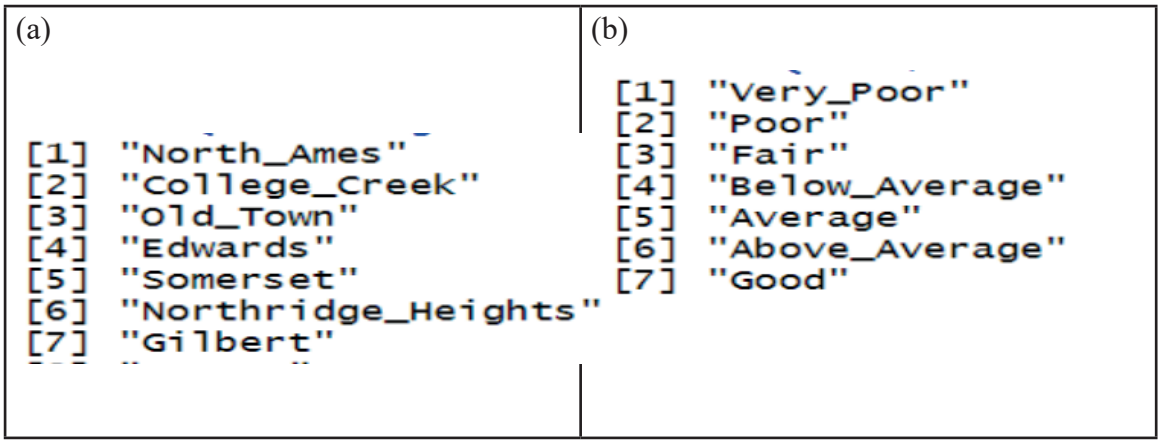
La codificación por etiquetas es una conversión numérica pura de los niveles de una variable categórica". Si una variable categórica es un factor y tiene niveles establecidos, entonces la conversión numérica será el orden de los niveles.

Sin embargo, hay que tener cuidado con la codificación por etiquetas en variables no ordenadas, ya que la mayoría de los modelos las tratarán como variables numéricas ordinales. Si una variable categórica está ordenada de forma natural, entonces esta codificación es una elección adecuada. En la siguiente figura, el valor de la codificación aparece entre corchetes.

Figura 9

Variable no apta para codificación por etiquetado(a) y variable apta para codificación por etiquetado(b)

Creación del autor Alfonso Prado



Codificación por One Hot para variables categóricas nominales

Esta codificación, es también conocida como variables "dummy" implementado en el paquete R caret. En primer lugar, a partir de una variable categórica X (que debe ser del tipo factor) con p niveles, se crean p variables nuevas y cada una de estas nuevas se asocia de forma individual a un elemento del conjunto Xi.

En segundo lugar, todas las variables nuevas toman el valor de 0, excepto aquella que representa el valor original del registro sin modificar, que tomará el valor 1. Siguiendo con nuestro dataset de ejemplo, consideramos la variable "Department"

Tabla 4

Codificación one-hot para variable categórica "Department"

Creación del autor Alfonso Prado

Department	Department0	Department1	Department2
Sales	1	0	0
R&D	0	1	0
HR	0	0	1

Codificación por One Cold para variables categóricas

Muy similar al anterior. La diferencia reside en que actúa de forma inversa: todas las variables son 1, excepto la asociada al valor que toma la variable categórica en cada observación.

El vector de codificación contiene las nuevas variables cuyos valores son todos 1, excepto aquel cuyo valor coincide con la variable original.

Codificación por Rank-Hot para variables ordinales

Hay casos en los que queremos indicarle al modelo que cierta etiqueta es mejor que otra etiqueta. Por ejemplo, en el dataset del ejemplo la variable Education (nivel de educación), en este caso utilizar One-Hot no es ideal ya que esto elimina la estructura de clasificación presente en los datos. Necesitamos una forma de transmitir la importancia, en este caso se pretende indicar que el nivel educación es de al menos x .

Tabla 5

Etiquetado por grado

Creación del autor Alfonso Prado

Education	Educación0	Educación1	Educación2	Educación3	Educación4
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1

Veamos el modelo completo

```
require(caret)

levels(data$MaritalStatus)

[1] "Divorced" "Married"  "Single"

dv <- dummyVars(~MaritalStatus,data=data)

trsf <- data.frame(predict(dv, newdata = data))

data_dv <- cbind(data,trsf)

#Creamos el modelo

glm(Attrition~MaritalStatus.Divorced +MaritalStatus.Single +

    MaritalStatus.Married , data=data_dv , family=binomial(link="logit"))

Call:  glm(formula = Attrition ~ MaritalStatus.Divorced + MaritalStatus.Single +
    MaritalStatus.Married, family = binomial(link = "logit"),
    data = data_dv)

Coefficients:
           (Intercept)  MaritalStatus.Divorced    MaritalStatus.Single
               -1.9235                 -0.2843                  0.8421
 MaritalStatus.Married
                    NA

Degrees of Freedom: 4299 Total (i.e. Null);  4297 Residual
Null Deviance:      3804
Residual Deviance: 3680         AIC: 3686
```

Referencias citadas en la Clase 13.

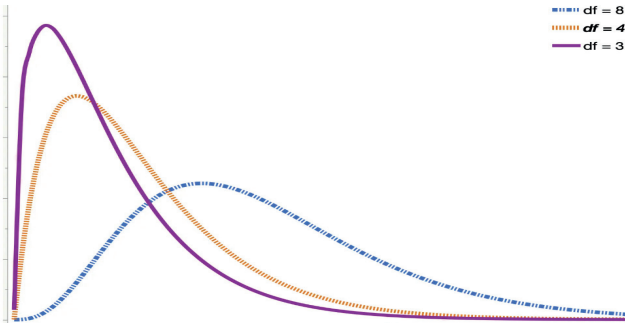
Szretter M. (2017) Apunte de Regresión Lineal, https://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

Rocha A.,(2020) Codificación de variables categóricas en aprendizaje automático, Universidad de Sevilla

Ramirez-Alan O,(2016), Chi-Cuadrado https://rpubs.com/osoramirez/111403

Mehdi Koosha, A. Amiri, (2019) The Effect of Link Function on the Monitoring of Logistic Regression Profiles, https://www.semanticscholar.org/

Definición de los términos citados en la Clase 13

Distribución Chi-cuadrado	La distribución ji cuadrado es una familia de distribuciones. Cada distribución se define por los grados de libertad. (Los grados de libertad se comentan en mayor detalle en las páginas sobre la prueba de bondad de ajuste y la prueba de independencia). En la siguiente figura se muestran tres distribuciones ji cuadrado diferentes, con distintos grados de libertad. 
Prueba de Chi cuadrado	La prueba Chi-cuadrado es una prueba de hipótesis utilizada para determinar si existe una relación entre dos variables categóricas. La prueba Chi-cuadrado comprueba si las frecuencias que se dan en la muestra difieren significativamente de las frecuencias que cabría esperar. Así, se comparan las frecuencias observadas con las esperadas y se examinan sus desviaciones.

Profundización Clase 1.

Conocer más de Modelo Logístico 1.

Conocer más de Modelos Logístico 2

Conocer más de Modelos logísticos 3

PUCE Virtual

La excelencia no se improvisa

síguenos