CLASIFICACIÓN: REGRESIÓN LOGÍSTICA

TRABAJO LABORATORIO (SESIÓN 2)

BOOK: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

An Introduction to Statistical Learning with Applications in R

Springer, 2013

Chapter 07

Regresión: Resumen Sesión anterior

- Hemos visto un conjunto de propiedades para determinar la bondad de la aproximación

- De la regresión no lineal, se han visto los comandos de R “lm()” y “gam()”. El primero corresponde a los conocidos como modelos no lineales linealizables. El segundo a los modelos puramente no lineales.

- Hemos visto 3 modelos de funciones polinomiales en R que se suelen utilizar para definir modelos no líneales: poly(), bs(), ns().

- Hemos comparado distintos modelos de regresión sobre distintos problemas. Se observa que aquellas aproximación con un elevado número de parámetros tiende al sobreajuste y pierden capacidad de predicción sobre ejemplos no vistos.

Clasificación: Regresión Logística

- Regresión logística
  - Regresión logística binaria
  - Regresión logística con más de 2 clases

Regresión Logística

Regresión logística binaria

- Haremos uso del *script3_Rlogistica.R*

- La regresión clásica trata de aproximar una función cuya variable se mueve en el espacio real.

- La regresión logística es una transformación de la regresión clásica para abordar problemas de clasificación, es decir, donde la variable de salida toma valores nominales o discretos.

- El proceso de transformación se ha explicado en la clase de teoría, aquí practicaremos con estos modelos.

Regresión Logística

Regresión logística binaria

- Ejecutamos el script hasta la línea 164 del scripts
- Esta es la base de datos con la que vamos a trabajar. Como podéis observar hay una modificación en la variable de salida.

Comando en R

```
datos <-data.frame(y=as.numeric(I(iris$Sepal.Width<3)),
                   x1=iris$Sepal.Length,
                   x2=iris$Petal.Length,
                   x3=iris$Petal.Width)
```

- El operador *I(x)* transforma una variable. En este caso, en verdadero o falso.
- El operador *as.numeric(x)* transforma en un valor numérico la variable x. En este caso, FALSE por un 0 y TRUE por un 1.
- Así, hemos lo transformamos en un problema de clasificación binaria.

Regresión Logística

Regresión logística binaria

- A continuación definimos los modelos con los que vamos a trabajar
- Los modelos *model0* y *model1* representan dos aproximadores lineales, el primero de ellos, el más habitual, y el segundo usando splines naturales.

Comando en R

```
model0 <-glm(y~.,data=datos)
model1 <-glm(y~ns(x1,4)+ns(x2,4)+ns(x3,4), data= datos)
model2 <-gam(y~ns(x1,4)+ns(x2,4)+ns(x3,4), family="binomial", data= datos)
model3 <-gam(y~ns(x1,16)+ns(x2,16)+ns(x3,16), family="binomial", data= datos)
model4 <-gam(y~ns(x1,4)*ns(x2,4)*s(x3,4), family="binomial", data= datos)
```

- Los otros tres modelos son puramente no lineales y hacen uso de splines suavizados.
- La principal diferencia en relación al uso previo que hemos hecho de ellos, es que aparece el parámetro family=“binomial”. Esto es lo que permite aplicarlos a clasificación.
- IMPORTANTE: los valores de ***y*** deben estar en el intervalo [0,1].

Regresión Logística

Regresión logística binaria

- Pasamos a evaluar el modelo 0

Comando en R

```
# Evaluacion del modelo 0
a <- Analisis(datos,model0)
AnalisisGrafico(datos,model0)
```

- Este se basa en glm() que ya habéis estudiado en el curso anterior. Aquí lo usaremos a modo de recuerdo.
- Aparecen 4 funciones definidas previamente en el script:
  - *Acierto(x,y)* devuelve el porcentaje de acierto del modelo. Se le pasa como primer argumento los valores reales y segundo argumento los datos predichos.
  - *ValidacionCruzada(d,n,m)* realiza una validación cruzada con *n* particiones sobre los datos *d* usando el modelo *m.*
  - *Analisis() y AnalisisGrafico() son versiones adaptadas de las que vimos en anteriores scripts.*

Regresión Logística

Regresión logística binaria

- Pasamos a evaluar el modelo 1

Comando en R

```
# Evaluacion del modelo 1
b <- Analisis(datos,model1)
AnalisisGrafico(datos,model1)
```

- Este se basa en glm() también, pero establece un modelo no lineal linealizable como ya hemos visto.

Regresión Logística

Regresión logística binaria

- Pasamos a evaluar el modelo 2,3 y 4 basados en GAM

Comando en R

```
# Evaluacion del modelo 2
c <- Analisis(datos,model2)
AnalisisGrafico(datos,model2)
```

Como se puede observar la estructura es semejante a los modelos que usan glm(). La diferencia está en type="response" para provocar una salida entre 0 y 1. El operador ***round*** redondea la salida para convertirla en valores binarios.

- Los modelos GAM de la biblioteca **mgcv**, al añadir la opción "family=binomial", devuelve una advertencia.

El fenómeno de separación o verosimilitud monótona se observa en el proceso de ajuste de un modelo logístico si la verosimilitud converge mientras al menos la estimación de un parámetro diverge hasta +/- el infinito

Regresión Logística
Regresión logística binaria

- Componemos en un data.frame los resultados obtenidos en el análisis.

Comando en R

```
# Comparacion entre los modelos
df <- data.frame(rbind(model0=a,model1=b,model2=c,
                       model3=d,model4=e),stringsAsFactors = FALSE)
df
```

- Los resultados obtenidos son los siguientes:

Resultado

	Acierto	CV
model0	0.76	0.747
model1	0.833	0.767
model2	0.847	0.787
model3	0.98	0.747
model4	1	0.34

Regresión Logística

Regresión logística binaria

	Resultado	
	Acierto	CV
model0	0.76	0.747
model1	0.833	0.767
model2	0.847	0.787
model3	0.98	0.747
model4	1	0.34

- Nos centraremos en Acierto y CV. El primero nos dice que capacidad de adaptación tiene sobre el conjunto de entrenamiento, mientras el segundo nos da la capacidad de predicción sobre ejemplos no vistos, usando una validación cruzada de 10.

- Podemos observar que model3 y model4 presentan sobreaprendizaje, mientras que los mejores modelos son model1 y model2, siendo model2 ligeramente mejor.

Regresión Logística

Regresión logística con más de 2 clases

- Los modelos glm() están adaptados para trabajar con más de dos clases. Para comprobarlo, vamos a trabajar con la bases de datos IRIS original, es decir, con 4 variables predictivas y una variable de clasificación con 3 clases.

Comando en R

```
datos <-data.frame(y=as.numeric(as.numeric(iris$Species)),
                   x4=iris$Sepal.Width,
                   x1=iris$Sepal.Length,
                   x2=iris$Petal.Length,
                   x3=iris$Petal.Width)
```

- Voy a transformar la variable de clase nominal en numérica.

Regresión Logística

Regresión logística con más de 2 clases

- Defino los mismos modelos glm() anteriores para este problema.

Comando en R

```
# modelo glm

model0 <- glm(y~.,data=datos)

model1 <-glm(y~ns(x1,4)+ns(x2,4)+ns(x3,4), data= datos)
```

- El model0 hace referencia a todas la variables predictivas y establece una aproximación lineal, mientras que el segundo ignora la variable x4 y utiliza un modelo no lineal linealizable usando splines naturales.

Regresión Logística

Regresión logística con más de 2 clases

- Aplico los procedimientos para valorar los dos modelos

Comando en R

```
# Evaluacion del modelo 0
a <- round(predict(model0, newdata = datos, type="response"))
Acierto(datos[,1],a)
ValidacionCruzada(datos,10,model0)
a <- Analisis(datos,model0)
AnalisisGrafico(datos,model0)
```

Resultado

	Acierto	CV
model0	0.973	0.96
model1	0.973	0.96

- Los dos modelos presenta una capacidad de predicción semejante.

Regresión Logística

Regresión logística con más de 2 clases

Ejercicio 2.1: El model1 no tenía en cuenta la variable x4. ¿Mejora el modelo si se incluye esta variable?

Redefine el model1 incluyendo x4 y compara ambos modelos.

Ejercicio 2.2: ¿Cuál es el mejor modelo que usa splines naturales de hasta grado 4 pero sólo hace uso de dos de las 4 variables predictivas?