**Ejercicio regresión logística**

Henry Navarro

1. **Credit-scoring**

Cuando los bancos quieren dar un cr´edito a un cliente, han de estimar primero el riesgo de impago por parte del mismo. El objetivo de los sistemas de credit-scoring es modelizar o predecir la probabilidad de pago o impago por parte de un cliente con cierto factores de riesgo. Los m´etodos de credit scoring se han convertido en una t´ecnica est´andar no s´olo para bancos, sino tambi´en para otras entidades financieras y grandes almacenes.

Uno de los primeros usos del credit scoring fue en el a´mbito de las tarjetas de cr´edito y de ah´ı se extendio a los cr´editos personales y a las hipotecas. La mayor´ıa de los problemas del credit scoring son m´as de tipo t´ecnico que te´orico. Para empezar. El primer problema se encuentra en los datos necesarios para el credit-scoring: deber´ıan estar disponibles la mayor cantidad posible de variables relevantes, pero el conseguir estas bases de datos es bastante caro, por lo que es necesario alcanzar un equilibrio entre el coste de los datos y los errores que se pueden cometer si la informaci´on de la que se dispone no es suficiente. Los bancos consiguen la informaci´on de sus propias fuentes internas (por ejemplo, de las solicitudes de cr´edito anteriores), de fuentes externas (cuestionarios y entrevistas personales) y de otras fuentes. Para conocer la situaci´on del solicitante, normalmente se recoge la siguiente informaci´on: edad, sexo, estado civil, nacionalidad, nivel educativo, nu´mero de hijos, salario, gastos, etc. Las variables que entran en el modelo utilizado para el credit-scoring deben ser elegidas con cuidado ya que la cantidad de datos puede ser elevada y computacionalmente compleja.

1. **Descripci´on de los datos**

Los datos corresponden a 800 clientes de un banco. Se dispone de 20 variables explicativas que pueden influir a la hora de conceder un cr´edito, de las cuales he seleccionado 7.

**Descripci´on de las variables**

Y:es una variable dicot´omica: ’bueno’, ’malo’que indica si el cliente se considera un buen o mal pagador.

*X*1**:** es una variable categ´orica con los niveles ‘no’, ‘cuenta.buena’, ‘cuenta.mala’, que corresponde al estado del cr´edito de la cuenta actual del cliente.

*X*2**:** variable num´erica que indica el nu´mero de meses de duraci´on del cr´edito.

*X*3**:** variable categ´orica con niveles ‘buen pagador previo’, ‘mal pagador previo’, si el cliente ha sido buen o mal pagador con anterioridad.

*X*4**:** variable categ´orica con niveles ‘personal’, ‘profesional’, indicando la intenci´on de uso del pr´estamo.

*X*5**:** variable num´erica que indica la cantidad del pr´estamo (en una cierta unidad mo- netaria).

*X*6**:** variable categ´orica con niveles ‘mujer’, ’hombre’, indicando el sexo del cliente

*X*7**:** variable categ´orica con niveles ‘solo’, nno ‘solo’, que indica si la persona tiene pareja o no.

La base de datos debe dividirse en dos partes. Un conjunto de entrenamiento y otro de test.

1. **Ejercicios**
   1. Ajusta 7 modelos de regresi´on log´ıstica (uno para cada una de las variables explicativas) y comenta los resultados.
   2. Ajusta el modelo con todas las variables. Utiliza el test de la raz´on de verosimilitud para determinar las variables que deben estar o no el modelo y compara los resultados lo que obtendría si utilizaras el criterio AIC.
   3. Basado en el modelo final, obtén intervalos de confianza al 90 % para el odds ratio de las variable que corresponden al tipo de uso del pr´estamo y al estado civil.
   4. Interpreta los coeficientes del modelo obtenido en t´erminos del odds ratio.
   5. Realiza la predicción para tu modelo final con los datos de entrenamiento.
   6. Calcula la matriz de confusión para los cutoff’s siguientes: 0.3, 0.5, 0.8. Analiza los resultados. ¿Cuál cutoff seleccionarías? ¿Por qué? Sugerencia: utiliza una función
   7. Grafica la curva ROC y analiza los resultados.
   8. Halla un valor de cutoff óptimo para los datos de entrenamiento y grafica la curva ROC. Analice muy bien los resultados, ¿qué observa para el TPR y TNR?
   9. Repita el ejercicio realizando una Cross-Validation, balanceando los datos y utilizando los datos de test para calcular todas las medidas de rendimiento. Para esto siga los pasos: a) balancee la muestra, b) cree un modelo con todas las variables, c) realice la predicción con el conjunto de test, d) Halle el valor de cutoff óptimo, e) aplique cross-validation.
   10. ¡A divertirse con Machine Learning! ¿Eres buen pagador? Utiliza el modelo anterior para determinar si eres elegible para un préstamo del monto que quieras, a pagar en el tiempo que desees. Analiza los resultados.
   11. Entrene el modelo con todos los datos. ¿Son los coeficientes estimados similares a los obtenidos s´olo con la muestra de entrenamiento?, ¿deberían serlo?