DEPARTAMENTO DE INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

EJERCICIO PRÁCTICO 14: REGRESIÓN LOGÍSTICA

CONTEXTO

Conocemos varias herramientas que facilitan la búsqueda y construcción de modelos de regresión lineal, además del proceso iterativo para conseguir un modelo confiable.

El objetivo de este ejercicio es utilizar herramientas y procedimientos análogos para crear y evaluar modelos de regresión logística (RLog) para predecir una variable dicotómica.

OBJETIVOS DE APRENDIZAJE

1. Preparar un conjunto de datos para la construcción de modelos RLog.
2. Iterar en el proceso de selección de variables, creación y evaluación de modelos RLog, hasta conseguir uno que sea confiable y satisfactorio.

ÉXITO DE LA ACTIVIDAD

El equipo es capaz de encontrar modelos RLog confiables y de buen desempeño al predecir una variable dependiente.

ACTIVIDADES

Para esta actividad usaremos los datos de medidas de atributos del vino que ya conocimos en el ejercicio práctico anterior.

1. El equipo copia el enunciado del problema asignado como comentarios de un script R.
2. El equipo lee el enunciado, descarga el archivo de datos (EP13 Datos.csv) desde UVirtual y selecciona las columnas para trabajar de acuerdo a las instrucciones.
3. El equipo construye los modelos solicitados usando la muestra correspondiente.
4. El equipo sube el script con las actividades anteriores comentando en detalle los pasos seguidos.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP14-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del sábado 8 de julio.

PREGUNTA (todos los grupos)

Ahora podemos construir un modelo de regresión logística para predecir la variable clase, de acuerdo con las siguientes instrucciones:

1. Definir la semilla a utilizar, que corresponde a los últimos cuatro dígitos del RUN (sin considerar el dígito verificador) del integrante de mayor edad del equipo.
2. Seleccionar una muestra de 120 vinos, asegurando que la mitad sean blancos y la otra mitad, tintos. Dividir esta muestra en dos conjuntos: los datos de 80 vinos (40 con clase “Blanco”) para utilizar en la construcción de los modelos y 40 vinos (20 con clase “Blanco”) para poder evaluarlos.
3. Seleccionar 6 variables predictoras de manera aleatoria (al igual que en el ejercicio anterior).

4. Seleccionar, de las otras variables, una que el equipo considere que podría ser útil para predecir la clase, justificando bien esta selección.
5. Usando el entorno R y paquetes estándares, construir un modelo de regresión logística con el predictor seleccionado en el paso anterior y utilizando de la muestra obtenida.
6. Agregue la variable seleccionada en el paso 4 al conjunto obtenido en el punto 3.
7. Usando herramientas estándares[1] para la exploración de modelos del entorno R, buscar entre dos y cinco predictores de entre las variables presentes en el conjunto obtenido en el paso anterior para construir un modelo de regresión logística múltiple.
8. Evaluar la confiabilidad de los modelos (i.e. que tengan un buen nivel de ajuste y son generalizables) y "arreglarlos" en caso de que tengan algún problema.
9. Usando herramientas del paquete caret, evaluar el poder predictivo de los modelos con los datos de los 40 vinos que no se incluyeron en su construcción en términos de sensibilidad y especificidad.

CRITERIOS DE EVALUACIÓN

- Seleccionan muestras de entrenamiento y prueba siguiendo las instrucciones dadas y asegurando que esta variable se encuentra balanceada en ellas.
- Seleccionan, justificando su utilidad de forma convincente, una variable de entre las no reservadas para explorar que utilizan para construir correctamente un modelo de RLogS para predecir la variable clase, evitando las variables obviamente correlacionadas (IMC, Peso, Estatura).
- Seleccionan un conjunto de variables relevantes para predecir la variable clase, utilizando correctamente gráficos y/o utilidades en paquetes de R para explorarlas, pero sin hacer uso del paquete caret, desde el conjunto de ocho variables elegidas aleatoriamente en el ejercicio pasado y respetando las otras restricciones indicadas en el enunciado.
- Construyen correctamente un modelo de RLogM para predecir la variable clase agregando las variables seleccionadas anteriormente al modelo RlogS que se tiene.
- Escriben comentarios y código en R correcto que verifica las condiciones que garantizan que tanto el modelo de RLogS como el modelo de RLogM obtenidos tienen un buen nivel de ajuste y son generalizables, interpretando explícita y correctamente los resultados obtenidos en cada paso y tomando acciones correctivas apropiadas de ser necesarias o comentando los riesgos asociados.
- Escriben código R correcto que evalúa la calidad predictiva tanto del modelo de RLogS como del modelo de RLogM obtenidos, en datos no utilizados para su construcción, y comparando correctamente los desempeños observados.
- Entregan conclusiones correctas y completas, basadas en las evaluaciones realizadas y el proceso de búsqueda de predictores seguido, respecto del modelo de RlogM obtenido.
- El script está completo, ordenado y bien indentado, logrando un programa que es fácil de seguir y que no requiere cambios para que funcione.
- El script está comentado paso a paso, con claridad (basta una lectura para entender) y con buena redacción y ortografía (<= 5 errores), usando vocabulario propio de la disciplina y el contexto del problema.

[1] Entenderemos esto como paquetes tradicionales, sin incluir el paquete `caret`.