Trabajo Práctico Final

Regresión Avanzada
Maestría en Explotación de Datos y Gestión del Conocimiento
Universidad Austral

Dra. Luciana Chiapella - Mg. Diego Marfetán Molina

Mayo 2024

Descripción del conjunto de datos

Trabajaremos con una base que contiene información sobre 777 universidades ubicadas en Estados Unidos. El conjunto de datos se encuentra disponible en el archivo `universidades.txt` y cuenta con las siguientes 9 variables:

1. `privada`: indica si la universidad es privada o no.
2. `aplicaciones`: cantidad de aplicaciones recibidas por la universidad durante el último año (cada estudiante que aspira a ingresar debe presentar una aplicación formal, a partir de la cual es admitido/a o rechazado/a), medida en miles de personas.
3. `ingresantes`: cantidad de aplicaciones aceptadas, medida en miles de personas.
4. `estudiantes`: cantidad total de estudiantes en carreras de grado, medida en miles de personas.
5. `top10`: porcentaje de ingresantes que fueron parte del 10% de estudiantes con mejores calificaciones en sus respectivas escuelas secundarias.
6. `cuota`: costo de la cuota de la universidad, medida en miles de dólares.
7. `prof_dr`: porcentaje de profesores de la universidad que poseen título de doctorado.
8. `razon`: tasa de estudiantes por profesor.
9. `tasa_grad`: porcentaje de estudiantes que se gradúan.

Consigna I: Regresión Lineal

1. Dividir aleatoriamente al conjunto de datos en bloques de entrenamiento (70%) y prueba (30%), definiendo una semilla para hacer que el resultado sea reproducible. Salvo que se exprese lo contrario, todas las consignas presentadas a continuación deben responderse empleando el conjunto de datos de entrenamiento.
2. Ajustar tres modelos diferentes de Regresión Lineal Múltiple con el método de los Mínimos Cuadrados Ordinarios (MCO), definiendo como variable respuesta a la tasa de graduación de cada universidad. Se debe justificar por qué se eligieron a esos tres modelos en particular (ejemplo: procesos automáticos, técnica del mejor subconjunto, criterios propios, etc.).
3. Comparar los 3 modelos a través de las siguientes métricas de performance: CME, PRESS, C_p, AIC y BIC. En base a los resultados observados, elegir un modelo "ganador".
4. Realizar un análisis de residuos sobre el modelo seleccionado en el punto anterior. Este análisis debe incluir el chequeo de cumplimiento de supuestos, presencia de colinealidad y casos atípicos y/o influyentes.
5. Consierando el modelo elegido, interpretar en palabras del problema los efectos estimados de los predictores sobre la respuesta, incluida su significación estadística (resultados del test t).

Consigna II: Regularización y Predicción

1. Ajustar el modelo elegido en la etapa anterior mediante la técnica Ridge, eligiendo el parámetro de penalidad mediante validación cruzada *k-fold*. Informar el valor óptimo de λ y comparar el resultado de este ajuste con el obtenido mediante MCO.

2. Ajustar el modelo elegido en la etapa anterior mediante la técnica Lasso, eligiendo el parámetro de penalidad mediante validación cruzada *k-fold*. Informar el valor óptimo de λ y comparar el resultado de este ajuste con el obtenido mediante MCO.
3. Evaluar la capacidad predictiva de los modelos MCO, Ridge y Lasso utilizándolos para estimar la tasa de graduación de universidades presentes en el conjunto de datos de prueba. Proveer alguna medida del error de predicción y determinar cuál de los tres modelos es el más adecuado.

Consigna III: Regresión Logística

1. Sobre el conjunto de datos original, definir la variable respuesta:

$$Y_i = \begin{cases} 0 & si \quad tasa\_grad_i < 0.75 \\ 1 & si \quad tasa\_grad_i \geq 0.75 \end{cases}$$

2. Dividir aleatoriamente al conjunto de datos inicial en bloques de entrenamiento (70%) y prueba (30%), definiendo una semilla para hacer que el resultado sea reproducible. Utilizar la función `createDataPartition()` del paquete `caret` para asegurarse que la proporción de éxitos en cada partición sea balanceada.
3. Ajustar un modelo de regresión logística para estudiar la variable binaria definida en el punto 1. Este modelo debe incluir todas las explicativas disponibles, a excepción de la variable `tasa_grad` original. En base al resultado obtenido, interpretar las razones de odds asociadas a predictores estadísticamente significativos al 5%.
4. Elegir el punto de corte óptimo para clasificación mediante el método de la curva ROC.
5. Utilizando el punto de corte hallado, clasificar las universidades del conjunto de datos de prueba y construir la matriz de confusión correspondiente. Informar e interpretar los valores observados de precisión, sensibilidad, especificidad, VPP, VPN, F_1 y κ.

Indicaciones

- Se deben entregar 2 archivos:
    1. Archivo en formato .pdf con **15 hojas de extensión como máximo** donde figuren las respuestas para cada una de las consignas. Si bien no es obligatorio, se recomienda el uso de R Markdown o Quarto para generar el documento.
    2. Archivo en formato .Rmd o .qmd que permita generar el reporte enviado, o bien un script de R con las sentencias utilizadas para el ajuste y análisis de los modelos.
- **Fecha de Entrega:** 19/06/2024 a través del aula virtual.
- **Equipos:** se permite trabajar en grupos de hasta 3 personas como máximo. **Importante:** Informar la conformación del grupo antes del 24/05/2024 por medio de un correo electrónico dirigido a ambos docentes (dmarfetan-ext@austral.edu.ar - lchiapella-ext@austral.edu.ar). Luego, solo uno de los integrantes deberá subir la resolución en el aula virtual.

¡¡¡Éxitos!!!