Tarea N°1

1EST17 - Aprendizaje Estadístico 1

Enver G. Tarazona

2022-10-22

El presente trabajo puede realizarse hasta en grupos de como máximo 3 personas. No olvide de indicar en la primera cara de su pdf los códigos e integrantes del grupo. Su trabajo de otro lado debe de ser reproducible; es decir, si se le pide un ejercicio práctico debe de hacerlo en R adjuntando su código y este debe de correr sin problemas en la corrección.

Problema 1 (14 puntos)

El conjunto de datos `Weekly` de la librería `ISLR2` muestra información sobre el rendimiento porcentual semanal para el índice bursátil S&P 500 entre 1990 y 2010. El objetivo principal del estudio es construir un modelo de clasificación que permita predecir si el mercado tuvo un rendimiento positivo o negativo en una semana determinada.

a) (1.0 puntos) Use el conjunto de datos completo para estimar un modelo logístico con `Direction` como respuesta y las cinco variables de retraso (`lag`) más `Volumen` como predictores. Use la función de resumen para imprimir los resultados de la estimación del modelo. ¿Alguno de los predictores parece ser estadísticamente significativo? ¿De ser asi, cuales?

b) (1.5 puntos) Considerando el modelo anteriormente estimado y la librería `caret`, muestre la matriz de confusión y las métricas necesarias para evaluar la eficiencia predictiva del modelo considerando un punto de corte de 0.5. Interprete todas las métricas halladas y concluya sobre los resultados obtenidos.

c) (1.5 puntos) Grafique la curva ROC y determine el área bajo la curva (AUC). Comente los resultados.

d) (2.0 puntos) Utilizando el criterio del índice J de Youden, determine cuál sería el punto de corte óptimo para un mejor balance entre la sensibilidad y especificidad del modelo. A partir de esto, muestre la nueva matriz de confusión y evalúe la eficiencia predictiva del modelo. Compare con los resultados obtenidos en (b). Resuma sus resultados en una tabla comparativa.

e) (2.0 puntos) Ahora ajuste el modelo de regresión logística utilizando como conjunto de datos de entrenamiento al periodo de 1990 a 2008, con`Lag2` como el único predictor (esquema de validación). Calcule la matriz de confusión y las métricas de evaluación del desempeño predictivo para los datos retenidos en el conjunto de validación (es decir, los datos de 2009 y 2010). Analice los resultados obtenidos.

f) (2.0 puntos) Compare el modelo estimado en la pregunta anterior con otros clasificadores (regresión binaria con enlaces probit, cloglog y cauchit, naive bayes con estimación no paramétrica de la densidad del predictor) y K-Vecinos más cercanos (encuentre el valor de K con validación cruzada). Encuentre aquel que tenga mejor desempeño para el conjunto de datos de validación. Utilice diversos criterios de comparación. Resuma sus resultados en una tabla comparativa.

g) (2.0 puntos) Presente **una sola gráfica comparativa** que muestre las curvas ROC para los clasificadores usados en la pregunta anterior. Analice los resultados presentados en la gráfica.

h) (2.0 puntos) Nuevamente teniendo a `Lag2` como el único predictor, realice la evaluación de todos los modelos de regresión binaria **implementando** (no usar una librería que brinde resultados automáticos como `caret`) el método de evaluación por validación cruzada usando $k = 10$ grupos (10-CV). Indique que modelo tiene mejor desempeño y compare con los resultados obtenidos con el esquema de validación.

Problema 2 (6 puntos)

Tema: Sesgo de selección y la "manera incorrecta de hacer CV".

Revise previamente el siguiente video sobre la forma correcta e incorrecta de realizar K-CV:

Cross-validation: right and wrong

La tarea aquí es diseñar un algoritmo para "probar" que el camino incorrecto es incorrecto y que el camino correcto es correcto.

a) ¿Cuáles son los pasos de dicho algoritmo? Escriba una sugerencia. Ayuda: ¿Cómo genera datos para predictores y etiquetas de clase, cómo realiza la tarea de clasificación, dónde se inserta el CV de forma correcta e incorrecta en su algoritmo? ¿Puedes hacer un dibujo esquemático de la forma correcta y la incorrecta?

b) Ahora estamos haciendo una simulación para ilustrar el problema del sesgo de selección en CV, cuando se aplica de manera incorrecta. Esto es lo que (conceptualmente) vamos a hacer:

Generar datos:

- Simular datos de alta dimensión ($p = 5000$ predictores) a partir de variables normales independientes o correlacionadas, pero con pocas muestras ($n = 50$).
- Asigna etiquetas de clase aleatoriamente (aquí solo 2). Esto significa que la "verdad" es que la tasa de clasificación errónea no puede ser muy pequeña. ¿Cuál es la tasa de clasificación errónea esperada (para este conjunto aleatorio)?

Tarea de clasificación:

- Elegimos algunos ($d = 25$) de los predictores (¿cómo? Simplemente seleccionamos aquellos con la correlación más alta con el resultado).
- Realizar una regla de clasificación sobre estos predictores.
- Luego ejecutamos CV ($k = 5$) solo en los predictores d (=manera incorrecta) o en todos los predictores $c + d$ (=manera correcta).
- Reportar errores de clasificación para ambas situaciones.

Una posible versión de esto se presenta en el código R a continuación. Revise el código y explique lo que se hace en cada paso, luego ejecute el código y observe si los resultados están de acuerdo con lo que esperaba. Realice cambios en el código R si desea probar diferentes estrategias.

Empezamos generando datos para $n = 50$ observaciones

```
library(boot)
# GENERAR DATOS; utilizar una semilla para la reproducibilidad
set.seed(4268)
n = 50  #número de observaciones
p = 5000  #número de predictores
```

```
d = 25  #principales predictores correlacionados elegidos

# generando datos para los predictores
xs = matrix(rnorm(n * p, 0, 4), ncol = p, nrow = n)  # forma simple de predictores no correlacionados
dim(xs)  # n times p
# generar etiquetas de clase independientes de los predictores, por
# lo que si todo se clasifica como clase 1, esperamos un 50% de
# errores en general
ys = c(rep(0, n/2), rep(1, n/2))  #Ahora realmente el 50% de cada una
table(ys)
```

CV INCORRECTO: Seleccione los 25 predictores más correlacionados fuera del CV.

```
corrs = apply(xs, 2, cor, y = ys)
hist(corrs)
selected = order(corrs^2, decreasing = TRUE)[1:d]  #top d correlaciones seleccionadas
data = data.frame(ys, xs[, selected])
```

Luego ejecute CV para el ajuste del clasificador: use la regresión logística e incorpore la función `cv.glm ()`

```
logfit = glm(ys ~ ., family = "binomial", data = data)
cost <- function(r, pi = 0) mean(abs(r - pi) > 0.5)
kfold = 10
cvres = cv.glm(data = data, cost = cost, glmfit = logfit, K = kfold)
cvres$delta
```

¡Observe una tasa de clasificación errónea cero!

CV CORRECTO: No preseleccionar predictores fuera del CV, sino como parte del CV. Necesitamos codificar esto nosotros mismos:

```
reorder = sample(1:n, replace = FALSE)
validclass = NULL
for (i in 1:kfold) {
    neach = n/kfold
    trainids = setdiff(1:n, (((i - 1) * neach + 1):(i * neach)))
    traindata = data.frame(xs[reorder[trainids], ], ys[reorder[trainids]])
    validdata = data.frame(xs[reorder[-trainids], ], ys[reorder[-trainids]])
    colnames(traindata) = colnames(validdata) = c(paste("X", 1:p), "y")
    foldcorrs = apply(traindata[, 1:p], 2, cor, y = traindata[, p + 1])
    selected = order(foldcorrs^2, decreasing = TRUE)[1:d]  #top d correlaciones seleccionadas
    data = traindata[, c(selected, p + 1)]
    trainlogfit = glm(y ~ ., family = "binomial", data = data)
    pred = plogis(predict.glm(trainlogfit, newdata = validdata[, selected]))
    validclass = c(validclass, ifelse(pred > 0.5, 1, 0))
}
table(ys[reorder], validclass)
1 - sum(diag(table(ys[reorder], validclass)))/n
```