Ejercicios: Selección de Variables, Regularización y Validación

Modelos Estadísticos de Predicción

AUTHOR
Víctor Aceña Gil - Isaac Martín de Diego

PUBLISHED
September 3, 2025

Ejercicio 1: Conceptual (Sobreajuste vs. Subajuste)

Explica con tus propias palabras qué es el sobreajuste (overfitting) y el subajuste (underfitting). Describe los síntomas de cada uno comparando el error de entrenamiento con el error de validación (o de test), y menciona la solución principal para cada problema.

Ejercicio 2: Práctico (Filtrado Básico)

Imagina que recibes un nuevo conjunto de datos con 50 predictores para un modelo de regresión. Antes de aplicar métodos computacionalmente costosos, decides hacer un filtrado inicial. Describe los cuatro criterios básicos que aplicarías para descartar variables de forma preliminar, según lo explicado en los apuntes. [cite: 1182-1190]

Ejercicio 3: Conceptual (AIC vs. BIC)

Tanto el AIC como el BIC son criterios para comparar modelos, pero se basan en filosofías distintas y tienen penalizaciones diferentes.

- a) Escribe la fórmula de la penalización por complejidad para el AIC y para el BIC.
- b) ¿Cuál de los dos criterios tenderá a seleccionar modelos más simples (más parsimoniosos)? ¿Por qué?
- c) Si tu objetivo principal es la precisión predictiva, ¿cuál de los dos criterios es generalmente preferido?

Ejercicio 4: Práctico (Best Subset y Criterios de Información)

Usa el conjunto de datos `mtcars` y la librería `leaps`.

- a) Utiliza la función `regsubsets()` para realizar una selección del mejor subconjunto (`best subset selection`) para predecir `mpg` usando el resto de variables.
- b) Obtén el `summary()` de los resultados. ¿Qué modelo (cuántas variables) es el mejor según el criterio Cp de Mallows?
- c) ¿Y cuál es el mejor modelo según el R^2 ajustado?
- d) ¿Coinciden ambos criterios en el número de variables del modelo óptimo?

Ejercicio 5: Conceptual (Métodos Stepwise)

Los métodos automáticos paso a paso (forward, backward, stepwise) son computacionalmente eficientes, pero el texto advierte sobre su uso. Menciona y explica brevemente tres de las principales limitaciones o problemas de estos métodos. [cite: 1317-1322]

Ejercicio 6: Práctico (Selección Backward Stepwise)

Utiliza el conjunto de datos `swiss` para predecir `Fertility`.

- a) Ajusta el modelo completo: `modelo_completo <- lm(Fertility ~ ., data = swiss)`.
- b) Utiliza la función `step()` para realizar una selección regresiva (backward) basada en el criterio AIC.
- c) Reporta la fórmula del modelo final que selecciona el algoritmo y su valor de AIC.

Ejercicio 7: Conceptual (Ridge vs. Lasso)

La regresión Ridge y Lasso son dos métodos de regularización muy populares, pero tienen un efecto fundamentalmente diferente sobre los coeficientes del modelo.

- a) ¿Qué tipo de penalización utiliza cada método (L_1 o L_2)?
- b) ¿Cuál de los dos métodos puede realizar selección de variables (es decir, anular coeficientes por completo)?
- c) Describe un escenario en el que preferirías usar Ridge sobre Lasso.

Ejercicio 8: Práctico (Regresión Lasso)

Utiliza el paquete `glmnet` y el conjunto de datos `mtcars` para predecir `mpg`.

- a) Prepara los datos: crea una matriz `x` para los predictores y un vector `y` para la respuesta.
- b) Utiliza la función `cv.glmnet()` para realizar una validación cruzada y encontrar el valor de `lambda` óptimo para una regresión Lasso (`alpha = 1`).
- c) Extrae y muestra los coeficientes del modelo Lasso ajustado con el `lambda.min`.
- d) ¿Qué variables ha eliminado el modelo (coeficientes iguales a cero)?

Ejercicio 9: Conceptual (Validación)

Explica la diferencia entre la estrategia de validación Train/Test Split simple y la Validación Cruzada k-fold. ¿Cuál es la principal ventaja de la validación cruzada sobre la división simple? ¿En qué situación (tamaño del dataset) recomendarías usar cada una?

Ejercicio 10: Práctico (Validación Cruzada)

Imagina que has ajustado dos modelos para predecir `mpg` en el dataset `mtcars`: 1. Un modelo simple: `mpg ~ wt + hp` 2. Un modelo complejo: `mpg ~ .` (todas las variables)

Utilizando la librería `caret` y la función `train()`, como se muestra en el `callout-tip` "La maldición del sobreajuste", configura y ejecuta una validación cruzada de 10 particiones para estimar el RMSE de ambos modelos. ¿Cuál de los dos modelos generaliza mejor a nuevos datos según esta estimación?