Ejercicios: Ingeniería de Características

Modelos Estadísticos de Predicción

AUTHOR
Víctor Aceña Gil - Isaac Martín de Diego

PUBLISHED
September 3, 2025

Ejercicio 1: Conceptual (Diagnóstico antes de Transformar)

El texto desaconseja fuertemente el enfoque de "ensayo y error" al aplicar transformaciones. Explica con tus propias palabras por qué la práctica de probar transformaciones hasta que mejore el R^2 es metodológicamente peligrosa. Menciona al menos tres de los riesgos específicos discutidos en los apuntes.

Ejercicio 2: Práctico (Escalado de Variables)

Utiliza el dataset `iris` de R y céntrate en las cuatro variables predictoras continuas (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`).

- a) Calcula la media y la desviación estándar de estas cuatro variables en su escala original. ¿Son sus escalas directamente comparables?
- b) Crea un nuevo data frame donde hayas aplicado la estandarización Z-Score a estas cuatro variables. Verifica que las nuevas variables tienen una media cercana a 0 y una desviación estándar de 1.
- c) ¿Por qué este paso de escalado es crucial antes de aplicar métodos de regularización como Ridge o Lasso, tal y como se menciona en el texto?

Ejercicio 3: Conceptual (Elección del Método de Escalado)

Describe un escenario hipotético para cada uno de los siguientes casos, explicando por qué el método de escalado elegido sería el más apropiado:

- a) Un escenario donde la estandarización Z-Score es preferible.
- b) Un escenario donde la normalización Min-Max es preferible.
- c) Un escenario donde el escalado robusto (usando mediana y IQR) es necesario.

Ejercicio 4: Práctico (Transformación para Linealizar)

En el tema anterior vimos que la relación en el dataset `cars` (entre `speed` y `dist`) no era perfectamente lineal.

- a) Ajusta el modelo `lm(dist ~ speed, data = cars)` y genera el gráfico de residuos vs. valores ajustados para confirmar visualmente la no linealidad (patrón curvo).

- b) Los apuntes sugieren que la transformación logarítmica es útil para relaciones con "rendimientos decrecientes". Propón y aplica una transformación (ej. sobre el predictor, la respuesta, o ambos) para intentar linealizar la relación. Por ejemplo, ajusta `lm(log(dist) ~ speed, data = cars)`.
- c) Genera de nuevo el gráfico de residuos vs. valores ajustados para el nuevo modelo. Compara ambos diagnósticos. ¿Ha mejorado la linealidad?

Ejercicio 5: Práctico (Transformación de Box-Cox)

Usa el dataset `Boston` de la librería `MASS`. La variable respuesta `medv` (valor mediano de la vivienda) es estrictamente positiva y tiene cierta asimetría.

- a) Carga la librería `MASS` y utiliza la función `boxcox()` para encontrar el valor de \(\lambda\) óptimo para la variable `medv` en un modelo simple frente a `lstat`. La fórmula sería `boxcox(medv ~ lstat, data = Boston)`.
- b) Observando el gráfico que se genera, ¿a qué valor "simple" (como -1, 0, 0.5, 1) se aproxima el \(\lambda\) óptimo?
- c) Basándote en este resultado, ¿cuál de las transformaciones clásicas (logarítmica, raíz cuadrada, inversa, etc.) sería la más recomendable para la variable `medv`?

Ejercicio 6: Conceptual (Codificación de Variables Categóricas)

Explica la diferencia fundamental entre la Codificación Ordinal y la Codificación One-Hot. Para cada una de las siguientes variables, indica qué método de codificación usarías y justifica tu elección:

- `mes`: ("Enero", "Febrero", "Marzo", ...)
- `nivel_riesgo`: ("Bajo", "Medio", "Alto", "Crítico")
- `pais_origen`: ("España", "Francia", "Alemania", "Italia")

Ejercicio 7: Práctico (Interacción entre Variables Continuas)

Usa el dataset `mtcars` para investigar si el efecto del peso de un coche (`wt`) sobre su consumo (`mpg`) depende de su potencia (`hp`).

- a) Ajusta un modelo que incluya un término de interacción entre `wt` y `hp`. Escribe la fórmula en R.
- b) Observa el `summary()` del modelo. ¿Es el término de interacción (`wt:hp`) estadísticamente significativo a un nivel de \(\alpha = 0.05\)?
- c) Basándote en el signo del coeficiente de la interacción, ¿cómo cambia el efecto del peso sobre el consumo a medida que aumenta la potencia? (Es decir, ¿el efecto negativo del peso se hace más fuerte o más débil en los coches más potentes?).

Ejercicio 8: Interpretación de una Interacción (Continua x Categórica)

Un investigador modela el salario (`salario`, en euros) en función de los años de experiencia (`experiencia`) y si el empleado tiene o no un máster (`master`, con "No" como categoría de referencia). El modelo ajustado es:

```
salario = 30000 + 1200*experiencia + 8000*masterSi + 300*experiencia:masterSi
```

- a) Escribe la ecuación de regresión específica para los empleados que no tienen un máster.
- b) Escribe la ecuación de regresión específica para los empleados que sí tienen un máster.
- c) Interpreta el coeficiente de la interacción (300). ¿Qué nos dice sobre el retorno económico de la experiencia para ambos grupos?

Ejercicio 9: Conceptual (Principio de Jerarquía)

Explica el principio de jerarquía en el contexto de los modelos de regresión con interacciones. Si un modelo incluye el término de interacción `A:B`, ¿por qué es una buena práctica incluir siempre los efectos principales `A` y `B`, incluso si sus tests t individuales no son significativos?

Ejercicio 10: Conceptual (Ingeniería de Características Avanzada)

Los apuntes discuten la creación de nuevas variables mediante ratios y combinaciones. Para cada uno de los siguientes escenarios, propón una nueva variable (feature) que podrías crear y explica qué relación podría capturar mejor que las variables originales por sí solas.

- a) Para predecir la rentabilidad de una tienda, tienes las variables `ventas_totales` y `numero_de_empleados`.
- b) Para predecir el riesgo de impago de un solicitante de préstamo, tienes las variables `ingresos_anuales` y `deuda_total`.