Selección de Variables, Regularización y Validación

Víctor Aceña - Isaac Martín

DSLAB

2025-09-10

El Problema de la Selección de Variables

En modelos de regresión con gran número de variables predictoras enfrentamos el desafío crítico de identificar qué variables son realmente relevantes

Los problemas principales:

- Inclusión de demasiadas variables $\rightarrow$ **sobreajuste**, pérdida de **interpretabilidad**, complejidad innecesaria
- Exclusión de variables importantes $\rightarrow$ **modelos subóptimos**
- Con p variables explicativas: 2^p modelos diferentes posibles
- Exploración exhaustiva **computacionalmente inviable** cuando p es grande

El objetivo del tema: Seleccionar el subconjunto óptimo de variables predictoras y validar la calidad del modelo resultante

Enfoques Principales del Tema

Seis enfoques sistemáticos:

1. **Filtrado basado en información básica**
   - Eliminación preliminar de variables irrelevantes
   - Criterios: variabilidad, correlación, VIF
2. **Criterios de bondad de ajuste**
   - Métricas para comparar modelos: AIC, BIC, Cp de Mallows
3. **Métodos de selección exhaustiva**
   - Evaluación sistemática: Best Subset Selection
4. **Métodos automáticos paso a paso**
   - Selección iterativa: forward, backward, stepwise
5. **Métodos basados en regularización**
   - Penalización de complejidad: Ridge, Lasso, Elastic Net
6. **Validación del modelo**
   - División train/test y validación cruzada

Proceso Completo de Construcción del Modelo

Etapas del proceso sistemático:

1. **Definición del problema y variables de interés**
2. **Recogida de datos** (fiabilidad, validez, ética, control de sesgos)
3. **Análisis Exploratorio de Datos (EDA)**
4. **Ajuste del modelo inicial**
5. **Evaluación del modelo** (R^2, ANOVA, significancia)
6. **Diagnóstico del modelo** (residuos, observaciones atípicas)
7. **Reducción de variables** $\leftarrow$ **Enfoque principal del tema**
8. **Validación del modelo** $\leftarrow$ **Enfoque principal del tema**

Este tema se centra en las etapas 7 y 8

Tipos de Experimentos para Recogida de Datos

Clasificación según el diseño:

- **Experimentos controlados:** Manipulación deliberada de variables independientes
- **Estudios observacionales exploratorios:** Sin intervención, registro natural
  - Transversales (un momento temporal)
  - Longitudinales (seguimiento temporal)
- **Estudios observacionales confirmatorios:** Testear hipótesis específicas
- **Encuestas y cuestionarios:** Datos estructurados sobre actitudes/comportamientos
- **Experimentos naturales:** Fenómenos naturales como intervención
- **Estudios de simulación:** Modelos matemáticos/computacionales
- **Datos secundarios:** Bases de datos existentes

Filtrado Basado en Información Básica

Objetivo: Filtrado preliminar antes de métodos sofisticados

Criterios de eliminación básicos:

1. **Variabilidad de las variables predictoras**

$$\text{Var}(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 < \epsilon$$

(típicamente $\epsilon = 0.01$)

2. **Correlación con la variable respuesta**

$$r_{X_j,Y} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Umbral mínimo: $|r_{X_j,Y}| > \delta$ (ej: $\delta = 0.1$)

Filtrado por Multicolinealidad

Detectar y eliminar variables redundantes:

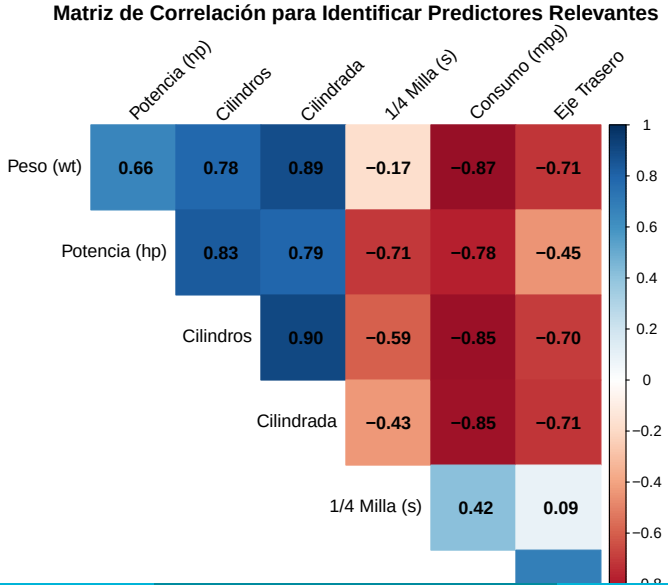
3. **Multicolinealidad extrema** Si $|r_{X_j,X_k}| > 0.95 \rightarrow$ eliminar una variable
4. **Factor de Inflación de la Varianza (VIF)**

$$VIF_j = \frac{1}{1 - R_j^2}$$

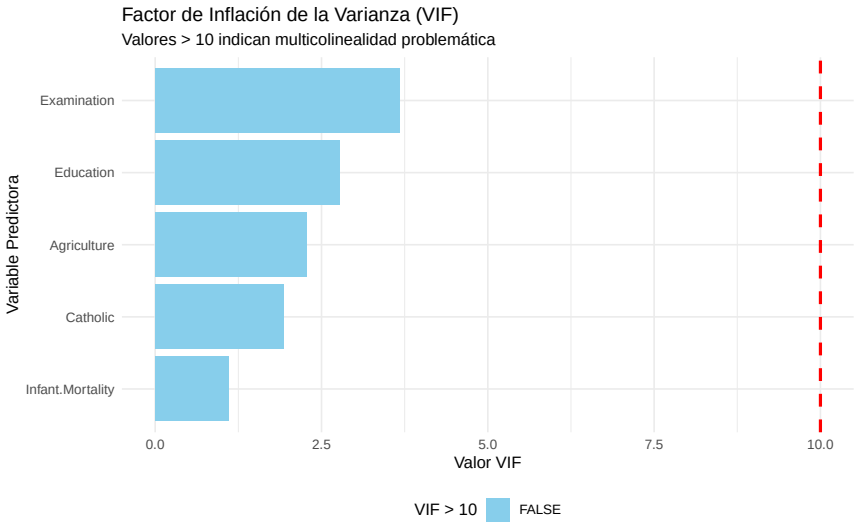
Valores $VIF_j > 10$ indican multicolinealidad problemática

Estrategia: Eliminar variables con mayor VIF iterativamente hasta que todos los VIF sean aceptables

Visualización: Matriz de Correlación



Visualización: Diagnóstico de Multicolinealidad



Criterios de Bondad de Ajuste

Problema: Equilibrar capacidad explicativa vs complejidad del modelo

Subajuste vs Sobreajuste: - Muy pocas variables $\rightarrow$ subajuste (underfitting) - Demasiadas variables $\rightarrow$ sobreajuste (overfitting)

Tres criterios principales:

1. **Criterio de Información de Akaike (AIC)**
2. **Criterio de Información Bayesiano (BIC)**
3. **Estadístico Cp de Mallows**

Estrategia: Seleccionar el modelo que minimice el criterio elegido

Criterio de Información de Akaike (AIC)

Fundamento: Teoría de la información de Hirotugu Akaike

Objetivo: Estimar la pérdida de información del modelo

Fórmula:

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2(p+1)$$

Componentes:

- $n \ln(SSE/n)$: **Bondad de ajuste** (relacionado con log-verosimilitud)
- $2(p+1)$: **Penalización por complejidad** (aumenta 2 unidades por parámetro)

Interpretación:

- Menor AIC = mejor modelo
- Asintóticamente eficiente
- **Orientado a predicción**

Criterio de Información Bayesiano (BIC)

Fundamento: Estadística bayesiana (Gideon Schwarz)

Objetivo: Encontrar el modelo más probable de ser el "verdadero"

Fórmula:

$$BIC = n \ln\left(\frac{SSE}{n}\right) + (p+1)\ln(n)$$

Diferencia clave con AIC:

- Penalización: $(p+1)\ln(n)$ en lugar de $2(p+1)$
- Más restrictivo cuando $n > 7$ (ya que $\ln(n) > 2$)

Características:

- **Consistencia:** Si el modelo verdadero está entre candidatos, P(selección) $\to 1$
- **Orientado a explicación**
- Favorece modelos más simples (parsimonia)

Estadístico Cp de Mallows

Fundamento: Error cuadrático medio de predicción

Objetivo: Modelo con bajo sesgo y baja varianza

Fórmula:

$$C_p = \frac{SSE_p}{MSE_{full}} - n + 2(p+1)$$

donde MSE_{full} es el error cuadrático medio del modelo completo

Interpretación:

- **Modelo bien especificado:** $C_p \approx p + 1$
- $C_p > p + 1$**:** modelo sesgado (variable importante omitida)
- $C_p \leq p + 1$**:** buen ajuste

Estrategia: Buscar modelos donde $C_p \approx p + 1$, elegir el menor entre ellos

¿Cuándo Usar Cada Criterio?

Si el objetivo principal es la predicción:

- **AIC** es la opción preferida
- Diseñado para minimizar error de predicción
- Penalización más moderada
- Ideal en contextos de pronóstico

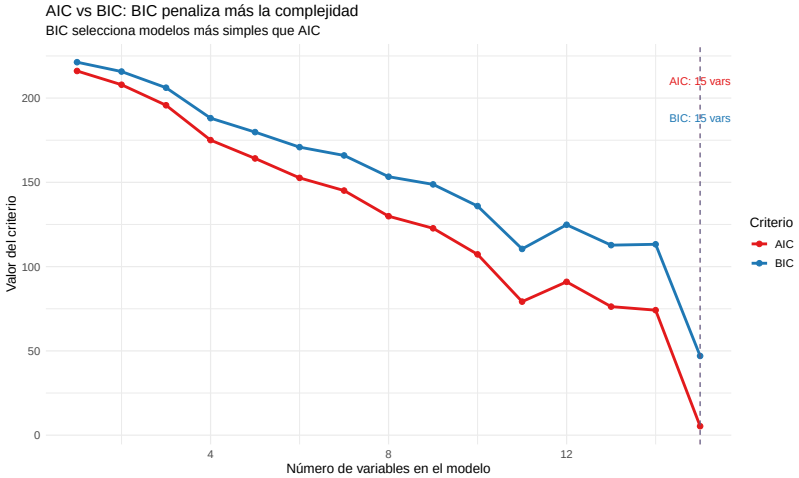
Si el objetivo es la explicación/inferencia:

- **BIC** es la elección más sólida
- Identifica el modelo más parsimonioso
- Penalización más fuerte contra sobreajuste
- Propiedad de consistencia en muestras grandes

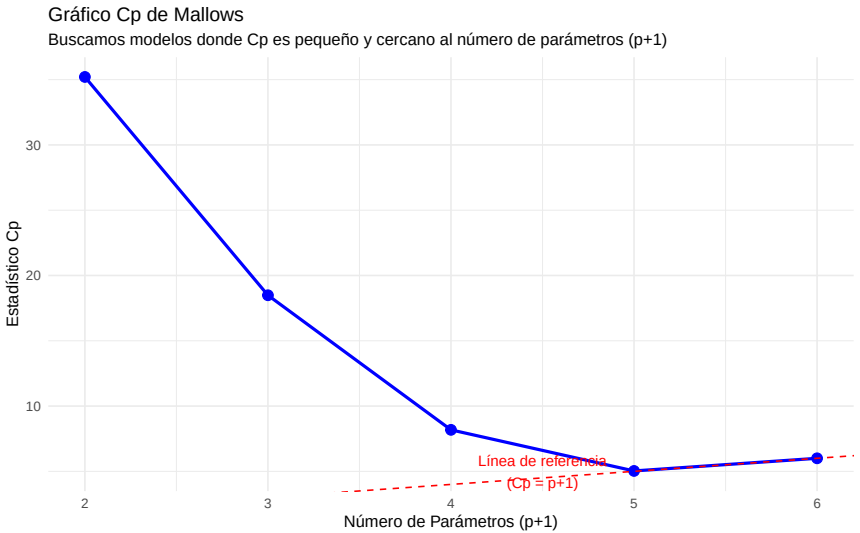
Para análisis exploratorio:

- **Cp de Mallows** es especialmente valioso
- Compromiso explícito entre sesgo y varianza
- Visualización clara del "codo" de complejidad óptima

Visualización: Criterios de Información (AIC vs BIC)



Visualización: Selección con Cp de Mallows



Métodos de Selección Exhaustiva

Best Subset Selection: Evalúa **todos** los subconjuntos posibles

Proceso:

- Para $k = 1, 2, ..., p$ variables
- Construir todos los modelos posibles con k variables
- Seleccionar el mejor modelo de cada tamaño según criterio elegido

Ventajas:

- Garantiza encontrar el modelo óptimo según el criterio
- Evaluación completa de todas las combinaciones
- Estándar para comparar otros métodos

Limitaciones:

- Complejidad computacional: 2^p modelos posibles
- Impracticable para $p > 15 - 20$
- Puede seleccionar modelos sobreajustados sin validación cruzada

Métodos Automáticos Paso a Paso

Principio: Construir modelo iterativamente, añadiendo o quitando predictores uno a uno

Forward Selection:

1. Comenzar con modelo nulo (solo intercepto)
2. Añadir variable que más mejora el criterio
3. Repetir hasta que ninguna variable mejore significativamente
4. **Problema:** No puede eliminar variables una vez incluidas

Backward Elimination:

1. Comenzar con modelo completo (todas las variables)
2. Eliminar variable menos significativa
3. Repetir hasta que todas las variables sean significativas
4. **Problema:** Requiere $n > p$

Métodos Automáticos Paso a Paso

Stepwise Regression:

1. Combina forward + backward
2. Puede añadir y eliminar variables
3. **Problema:** Solo encuentra óptimo local

Limitaciones importantes de métodos automáticos:

- **Inestabilidad:** Pequeños cambios en datos pueden alterar el modelo
- **Invalidez de p-valores:** Múltiples comparaciones sesgan la inferencia
- **Óptimo local:** No garantizan la mejor combinación
- **Inflación del error tipo I:** Sin corrección para comparaciones múltiples

Uso recomendado: Como herramientas exploratorias, no para inferencia final

Métodos Basados en Regularización

Principio: Introducir penalización en la función de ajuste del modelo

Objetivos:

- Controlar el sobreajuste reduciendo complejidad
- Forzar selección de subconjunto más parsimonioso
- Mejorar estabilidad y precisión del modelo

Tres métodos principales:

- **Ridge Regression:** Penalización $L_2 = \lambda \sum \beta_j^2$
- **Lasso:** Penalización $L_1 = \lambda \sum |\beta_j|$
- **Elastic Net:** Combina $L_1 + L_2$

Ventaja clave: Control automático del balance sesgo-varianza

Ridge Regression

Fundamento: Penalización L_2 en la estimación de coeficientes

Formulación:

$$SSE_{ridge} = \|\mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Estimación:

$$\widehat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\,\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Interpretación del parámetro λ:

- $\lambda = 0$: equivalente a regresión lineal tradicional (OLS)
- λ aumenta: coeficientes se reducen en magnitud
- λ muy grande: coeficientes se acercan a cero

Propiedades:

- Manejo de multicolinealidad

Regresión Lasso

Fundamento: Penalización L_1 que permite eliminación de variables

Formulación:

$$SSE_{\text{lasso}} = \|\mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Diferencia clave con Ridge:

- Ridge reduce magnitud de coeficientes
- **Lasso puede eliminar variables por completo** (coeficientes $= 0$)

Interpretación del parámetro λ:

- $\lambda = 0$: regresión lineal tradicional
- λ aumenta: más coeficientes $\to 0$
- λ muy grande: elimina demasiadas variables

Propiedades:

- Selección automática de variables
- Manejo de multicolinealidad
- Simplicidad e interpretabilidad
- Reduce sobreajuste

Elastic Net: Combinando lo Mejor de Ridge y Lasso

Fundamento: Combinación de penalizaciones Ridge (L_2) y Lasso (L_1)

Formulación:

$$SSE_{\text{Elastic Net}} = \|\mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right]$$

Parámetro α controla la mezcla:

- $\alpha = 1 \rightarrow$ Comportamiento como Lasso
- $\alpha = 0 \rightarrow$ Comportamiento como Ridge
- $0 < \alpha < 1 \rightarrow$ Combinación de ambos métodos

Ventajas principales:

- **Manejo superior de multicolinealidad**
- **Selección de variables más estable**
- **Evita selección arbitraria cuando hay grupos correlacionados**

Cuándo Usar Cada Método de Regularización

Ridge Regression:

- Todas las variables aportan información
- Fuerte multicolinealidad presente
- Objetivo: reducir varianza sin eliminar variables

Lasso:

- Muchas variables irrelevantes esperadas
- Selección sparse deseable
- Interpretabilidad prioritaria

Elastic Net:

- Variables correlacionadas en grupos
- Balance entre selección y estabilidad
- Rendimiento predictivo como objetivo principal

Estrategia práctica: Optimizar α mediante validación cruzada junto con λ

Selección del Parámetro de Regularización

El problema: ¿Cómo elegir el valor óptimo de lambda?

La solución: Validación cruzada

Proceso:

1. Definir secuencia de valores lambda candidatos
2. Para cada lambda, calcular error de validación cruzada
3. Seleccionar lambda que minimiza el error

Dos criterios principales:

- **lambda_min:** Valor que minimiza el error de CV
- **lambda_1SE:** Valor más grande cuyo error está dentro de 1 error estándar del mínimo

Regla 1-SE: Preferir modelo más simple (mayor lambda) si su error es comparable al mínimo

Estrategias de Validación: Visión General

Partición inicial (paso obligatorio):

Antes de cualquier análisis, dividir datos originales en:

1. **Datos de modelado (80%):** Para todo el proceso de construcción
2. **Conjunto de prueba final (20%):** Guardado para evaluación final

Dentro de los datos de modelado, tres estrategias principales:

1. **División Train/Test simple**
2. **Validación cruzada k-fold**
3. **Leave-One-Out Cross-Validation (LOOCV)**

Cada estrategia tiene sus ventajas e inconvenientes según el contexto del problema

División Train/Test Simple

Concepto: División única de los datos de modelado

Proceso:

- **Conjunto de entrenamiento (70-80%):** Ajustar el modelo
- **Conjunto de test (20-30%):** Evaluar rendimiento

Ventajas:

- Computacionalmente muy eficiente
- Fácil de implementar y entender
- Apropiado para datasets grandes

Desventajas:

- **Alta variabilidad:** Resultados dependen de la división específica
- Puede ser optimista o pesimista según qué observaciones caigan en test
- Menos datos disponibles para entrenamiento

Cuándo usar: Datasets grandes ($n > 1000$), recursos limitados, evaluación rápida

Validación Cruzada k-fold

Concepto: Múltiples evaluaciones para obtener estimación más estable

Proceso de k-fold CV:

1. Dividir datos en k particiones de tamaño similar
2. Para cada partición $i = 1, 2, ..., k$:
   - Usar partición i como conjunto de test
   - Usar las $k - 1$ particiones restantes como entrenamiento
   - Calcular métrica de error
3. **Error de CV:** $CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{Error}_i$

Valores típicos: $k = 5$ o $k = 10$

Ventajas:

- Estimación más estable y menos sesgada
- Todos los datos se usan para entrenamiento y test
- Reduce variabilidad de la estimación

Leave-One-Out Cross-Validation (LOOCV)

Concepto: Caso extremo donde $k = n$ (número de observaciones)

Proceso:

- Para cada observación i:
  - Entrenar modelo con $n - 1$ observaciones
  - Predecir la observación i excluida
  - Calcular error de predicción
- **Error LOOCV:** $CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} (\frac{y_i - \hat{y}_i}{1 - h_{ii}})^2$

Ventajas:

- Estimación prácticamente insesgada
- Determinística (no depende de divisiones aleatorias)
- Máximo uso de datos para entrenamiento

Desventajas:

- Computacionalmente costoso

Cuándo Usar Cada Estrategia de Validación

Train/Test Split:

- Dataset grande ($n > 1000$)
- Recursos computacionales limitados
- Necesidad de evaluación rápida
- Primera aproximación al problema

Validación Cruzada k-fold:

- Dataset de tamaño moderado ($100 < n < 1000$)
- Balance entre eficiencia y precisión
- Estimación robusta del rendimiento
- **Más recomendado en general**

LOOCV:

- Dataset pequeño ($n < 100$)
- Necesidad de estimación menos sesgada
- Recursos computacionales abundantes

Métricas de Rendimiento en Validación

Una vez obtenidas las predicciones, necesitamos "calificar" el modelo

Raíz del Error Cuadrático Medio (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Características del RMSE:

- Penaliza desproporcionadamente errores grandes
- Sensible a valores atípicos
- Mismas unidades que la variable respuesta
- Interpretación: "desviación típica de los residuos"

Métricas de Rendimiento en Validación

Error Absoluto Medio:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Características del MAE:

- Trata todos los errores proporcionalmente
- Más robusto frente a valores atípicos
- Interpretación directa: "error promedio en valor absoluto"

¿Cuándo usar cada métrica?

- **RMSE:** Cuando errores grandes son especialmente problemáticos
- **MAE:** Cuando se prefiere robustez frente a valores atípicos
- **Ambas:** Para análisis completo del rendimiento predictivo

Interpretación de Errores: Diagnóstico del Ajuste

La comparación clave: Error en entrenamiento vs Error en validación

Sobreajuste (Overfitting):

- **Síntoma:** Error entrenamiento bajo + Error validación mucho más alto
- **Causa:** Modelo memoriza ruido específico de los datos de entrenamiento
- **Solución:** Simplificar modelo, usar regularización, más datos

Subajuste (Underfitting):

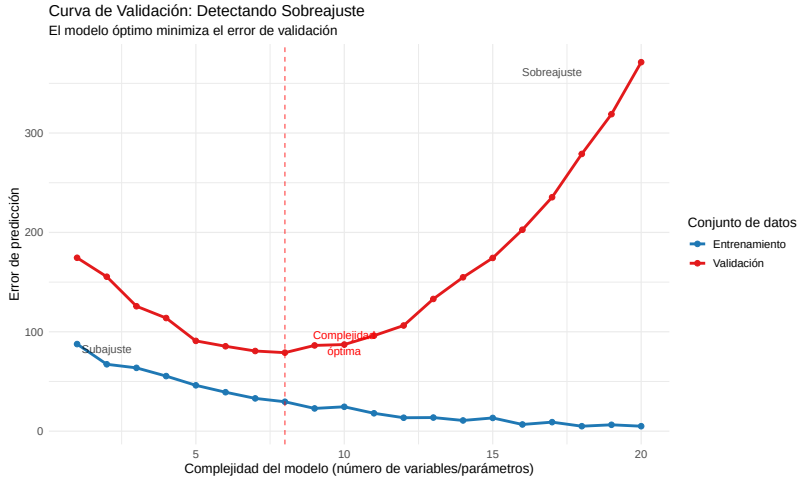
- **Síntoma:** Error entrenamiento alto + Error validación alto y similar
- **Causa:** Modelo demasiado simple, no captura estructura subyacente
- **Solución:** Aumentar complejidad, añadir variables, términos de interacción

Modelo bien calibrado:

- **Síntoma:** Error entrenamiento y validación similares y bajos
- **Interpretación:** Buen equilibrio entre sesgo y varianza

Resultado esperado: El modelo correcto tendrá mejor rendimiento en validación cruzada

Error de Entrenamiento vs Validación



El Conjunto de Prueba Final: La Evaluación Definitiva

Después de todo el proceso de modelado:

1. Filtrado de variables
2. Selección del mejor método
3. Optimización de hiperparámetros
4. Validación cruzada para elegir modelo final

El paso final: Evaluar el modelo seleccionado en el conjunto de prueba final

¿Por qué es necesario?

- La validación cruzada se usó para **tomar decisiones** sobre el modelo
- Existe riesgo de sobreajuste al proceso de validación mismo
- Necesitamos una evaluación completamente independiente

Interpretación:

- Error similar a validación cruzada $\rightarrow$ modelo robusto
- Error mucho mayor $\rightarrow$ posible sobreajuste al proceso de modelado

Advertencias Importantes

Los métodos stepwise (forward, backward, stepwise) requieren precaución especial

Problemas fundamentales:

1. **Invalidez de p-valores:** Los p-valores y errores estándar están sesgados
2. **Inestabilidad:** Pequeños cambios en datos pueden cambiar radicalmente el modelo
3. **Óptimo local:** No garantizan encontrar la mejor combinación de variables
4. **Inflación del error tipo I:** Múltiples comparaciones sin corrección

Uso recomendado:

- Como herramientas **exploratorias** únicamente
- Generar modelos candidatos para evaluación posterior
- Siempre validar con técnicas robustas
- No reportar p-valores del modelo final como definitivos

Proceso Completo de Selección y Validación

Flujo de trabajo recomendado:

1. **Partición inicial:** Separar conjunto de prueba final (20%)
2. **En datos de modelado (80%):**
   - Filtrado básico de variables
   - Aplicar métodos de selección (exhaustivos, stepwise, regularización)
   - Comparar modelos con validación cruzada
   - Seleccionar modelo final
3. **Evaluación final:** Probar modelo seleccionado en conjunto de prueba
4. **Reportar:** Error de validación cruzada Y error en conjunto de prueba

Criterios de decisión:

- Número de variables vs tamaño de muestra $\rightarrow$ método de selección
- Objetivo (predicción vs explicación) $\rightarrow$ criterio de información
- Multicolinealidad $\rightarrow$ regularización vs selección clásica

Consideraciones Prácticas

Antes del modelado:

- EDA completo para entender los datos
- Conocimiento del dominio para variables importantes
- Objetivo claro: ¿predicción o explicación?
- Relación entre tamaño muestral y número de variables

Durante la selección:

- Usar validación cruzada para todos los hiperparámetros
- Comparar múltiples métodos de selección
- No guiarse solo por métricas: considerar interpretabilidad
- Documentar todas las decisiones tomadas

Después de la selección:

- Diagnóstico completo de residuos del modelo final
- Análisis de sensibilidad a observaciones influyentes
- Intervalos de confianza para coeficientes importantes

Conclusiones

Lo aprendido en este tema:

1. **Filtrado inicial:** Elimina problemas básicos de forma eficiente
2. **Criterios de información:** Guían comparación objetiva de modelos
3. **Métodos exhaustivos:** Garantizan óptimo pero son computacionalmente costosos
4. **Regularización:** Controla sobreajuste y realiza selección automáticamente
5. **Validación:** Indispensable para evaluar capacidad de generalización

Recomendaciones principales:

- **Combinar métodos:** Ningún método es perfecto en todas las situaciones
- **Validar siempre:** Con datos que el modelo no ha visto
- **Preferir simplicidad:** Cuando el rendimiento es comparable
- **Incorporar conocimiento del dominio:** Los datos no lo dicen todo

El mejor modelo es aquel que resuelve el problema con la mayor simplicidad.