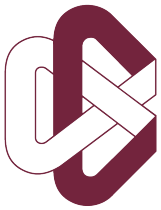


CIMAT

Centro de Investigación en Matemáticas
Unidad Monterrey

Análisis Multimodal

Tarea 2

Gustavo Hernández Angeles

16 de noviembre de 2025

Índice

Ejercicio 1:

Utilizando el conjunto de datos `College` disponible en la libreria `ISLR`, predice el número de solicitudes recibidas (`Apps`) utilizando las otras variables del conjunto de datos.

a) Divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.

b) Ajusta un modelo lineal utilizando mínimos cuadrados en el conjunto de entrenamiento y reporta el error de prueba obtenido.

c) Ajusta un modelo de regresión ridge en el conjunto de entrenamiento, con λ elegido por validación cruzada. Reporta el error de prueba obtenido.

d) Ajusta un modelo Lasso en el conjunto de entrenamiento, con λ elegido por validación cruzada. Reporta el error de prueba obtenido, junto con el número de estimaciones de coeficientes distintos de cero.

e) Ajusta un modelo PCR en el conjunto de entrenamiento, con M elegido por validación cruzada. Reporta el error de prueba obtenido, junto con el valor de M seleccionado por validación cruzada.

f) Ajusta un modelo PLS en el conjunto de entrenamiento, con M elegido por validación cruzada. Reporta el error de prueba obtenido, junto con el valor de M seleccionado por validación cruzada.

g) Comenta los resultados obtenidos. ¿Con qué precisión podemos predecir la cantidad de solicitudes universitarias recibidas? ¿Hay mucha diferencia entre los errores de prueba resultantes de estos cinco enfoques?

h) Propón un modelo (o un conjunto de modelos) que parezca funcionar bien en este conjunto de datos y justifica tu respuesta. Asegúrate de evaluar el rendimiento del modelo utilizando el error del conjunto de validación, la validación cruzada o alguna otra alternativa razonable, en lugar de utilizar el error de entrenamiento. ¿El modelo que elegiste incluye todas las características del conjunto de datos? ¿Por qué o por qué no?

1.1 Inciso a)

Se utilizó el conjunto de datos `College` y, para evitar fuga de información, se eliminaron las variables `Accept` y `Enroll` del análisis, ya que están determinadas después de `Apps` o están fuertemente condicionadas por ella; incluirlas inflaría artificialmente el desempeño en prueba. Posteriormente, se dividió el conjunto en entrenamiento y prueba usando un 50 % para cada uno (muestra aleatoria con semilla fija).

1.2 Inciso b)

Se ajustó un modelo lineal utilizando Mínimos Cuadrados Ordinarios (MCO) en el conjunto de entrenamiento, empleando todos los predictores. Las variables que resultaron ser estadísticamente significativas (con $p < 0.1$) en dicho modelo fueron:

- `F.Undergrad` $(***)$

- `Room.Board` (***)
- `Expend` (**)
- `Grad.Rate` (*)
- `perc.alumni` (.)

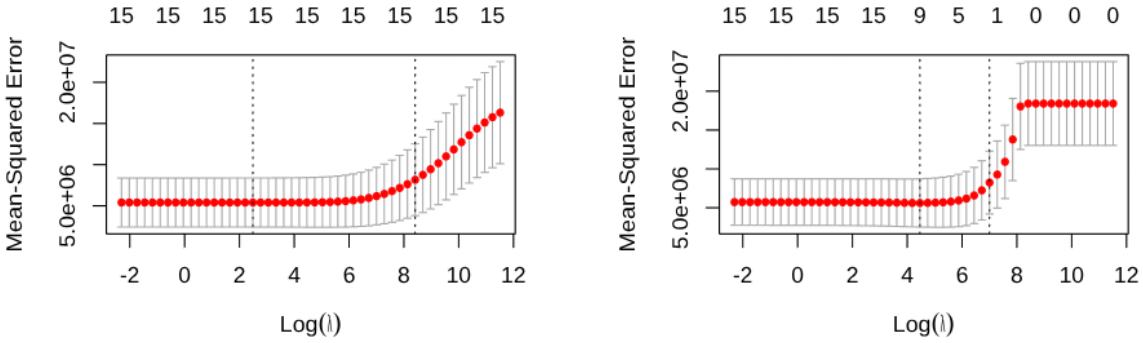
mientras que las demás variables no mostraron significancia estadística en el modelo ajustado. Al evaluar el rendimiento de este modelo en el conjunto de prueba, se obtuvo un Error Cuadrático Medio (MSE) de **2,551,734**. Esto corresponde a un Error Cuadrático Medio Raíz (RMSE) de **1,597.4**.

1.3 Inciso c)

Se ajustó una regresión *ridge* con validación cruzada de 5 pliegues para seleccionar λ. El valor $\lambda_{\text{mín}}$ elegido por CV fue **12.07** (y $\lambda_{1se} =$ **4498.43** como alternativa más parsimoniosa). Con $\lambda_{\text{mín}}$, el desempeño en prueba fue: MSE **2,530,947** (RMSE **1,590.9**). Con λ_{1se} el error aumentó a MSE **3,552,978** (RMSE **1,884.9**).

1.4 Inciso d)

Se ajustó un modelo *lasso* también con validación cruzada de 5 pliegues. Se obtuvo $\lambda_{\text{mín}} =$ **86.85** y $\lambda_{1se} =$ **1098.54**. Con $\lambda_{\text{mín}}$, el error de prueba fue MSE **2,395,665** (RMSE **1,547.8**), y con λ_{1se} el MSE fue **3,438,634**. El modelo con $\lambda_{\text{mín}}$ seleccionó **9** coeficientes distintos de cero: `Top10perc`, `Top25perc`, `F.Undergrad`, `Room.Board`, `Personal`, `S.F.Ratio`, `perc.alumni`, `Expend` y `Grad.Rate`.



(a) Selección de λ por validación cruzada – Ridge. (b) Selección de λ por validación cruzada – Lasso.

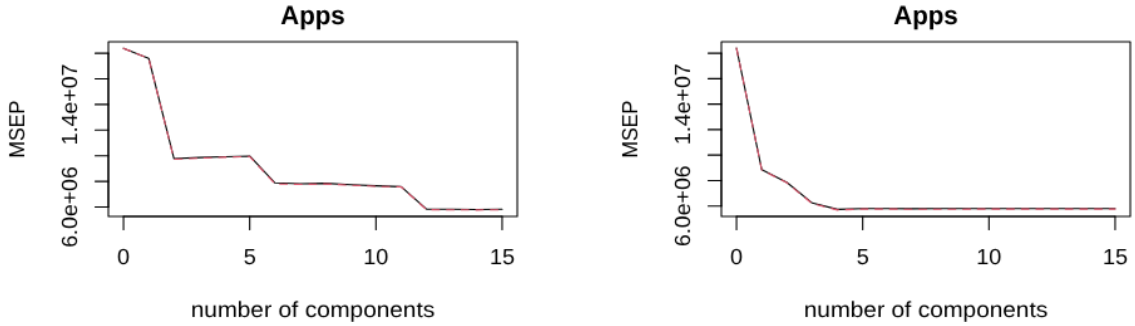
Figura 1.1: Comparación de curvas de validación cruzada para la selección de λ.

1.5 Inciso e)

Para PCR (componentes principales como regresores), con escalamiento y validación cruzada, el menor MSEP se obtuvo con $M =$ **12** componentes. En prueba, el desempeño fue MSE **2,389,249** (RMSE **1,545.7**). Nótese que con 6 componentes se explica aproximadamente el 80.8 % de la varianza en $\mathbf{X}$ y 61.7 % de `Apps`; con 9 componentes, 91.1 % en $\mathbf{X}$ y 62.7 % en `Apps`.

1.6 Inciso f)

Para PLSR (componentes latentes que maximizan covarianza), la validación cruzada sugirió $M = \mathbf{4}$ componentes. En el conjunto de prueba se obtuvo MSE **2,457,676** (RMSE **1,567.7**).



(a) Curva de MSEP para PCR.

(b) Curva de MSEP para PLSR.

Figura 1.2: Selección del número de componentes M mediante validación cruzada.

1.7 Inciso g)

En términos de precisión, todos los métodos logran errores de prueba de magnitud similar: RMSE entre $\approx$ **1,545** y **1,600** solicitudes. Dado que el promedio de `Apps` ronda $\sim$**3,000**, el error típico es del orden de la mitad del promedio, lo que indica capacidad predictiva moderada pero con variabilidad sustancial no explicada. Entre los enfoques, **PCR (M=12)** y **Lasso** con $\lambda_{\text{mín}}$ fueron los mejores (MSE $\approx 2.39 \times 10^6$), seguidos de **Ridge** (ligeramente peor) y **PLSR**. El modelo lineal MCO quedó rezagado respecto a PCR/Lasso, aunque no por un margen muy grande.

1.8 Inciso h)

Una propuesta razonable es utilizar **Lasso con** $\lambda_{\text{mín}}$: ofrece un desempeño competitivo (prácticamente el mejor MSE) y además un modelo *parco* con solo 9 predictores, lo que facilita interpretación y despliegue. Usar λ_{1se} reduciría aún más la complejidad, pero en este caso incurre en una pérdida de precisión considerable. Como alternativa si la interpretabilidad de coeficientes es secundaria, **PCR con** $M = 12$ proporciona un error prácticamente indistinguible del de Lasso. En cualquier caso, se recomienda evaluar con validación cruzada repetida o un esquema de particiones múltiples para estabilizar las estimaciones de error.

Finalmente, el modelo propuesto excluye deliberadamente `Accept` y `Enroll` por tratarse de variables que pueden introducir fuga de información. El conjunto de covariables retenido por Lasso incluye factores plausibles desde el punto de vista sustantivo (tamaño de matrícula, cuotas, gasto institucional y tasa de graduación), lo que respalda su uso en la práctica.

Ejercicio 2:

Es bien sabido que la regresión ridge tiende a dar valores de coeficientes similares a las variables correlacionadas, mientras que lasso puede dar valores de coeficientes totalmente diferentes a las variables correlacionadas. Se explorará esta propiedad en un entorno sencillo.

Supongamos que $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Además, supongamos que $y_1 + y_2 = 0$ y $x_{12} + x_{22} = 0$, de modo que la estimación del intercepto en mínimos cuadrados, regresión de Ridge o en el modelo de lasso es cero: $\hat{\gamma}_0 = 0$.

a) Plantea el problema de la optimización con la regresión ridge bajo estas suposiciones.

b) Argumenta que bajo estas suposiciones, las estimaciones de los coeficientes de ridge satisfacen $\hat{\beta}_1 = \hat{\beta}_2$.

c) Plantea el problema de la optimización con la regresión lasso bajo estas suposiciones.

d) Argumenta que en este contexto, los coeficientes de lasso $\hat{\beta}_1$ y $\hat{\beta}_2$ no son únicos; es decir, hay muchas soluciones posibles al problema de optimización en (c). Describe estas soluciones.