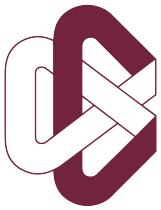


CIMAT

Centro de Investigación en Matemáticas
Unidad Monterrey

Análisis Multimodal

Tarea 2

Gustavo Hernández Angeles

18 de noviembre de 2025

Índice

1 Ejercicio 1: **3**
1.1 Inciso a) . . . 3
1.2 Inciso b) . . . 3
1.3 Inciso c) . . . 4
1.4 Inciso d) . . . 4
1.5 Inciso e) . . . 4
1.6 Inciso f) . . . 5
1.7 Inciso g) . . . 5
1.8 Inciso h) . . . 6

2 Ejercicio 2: **7**
2.1 Inciso a) . . . 8
2.2 Inciso b) . . . 8
2.3 Inciso c) . . . 9
2.4 Inciso d) . . . 9

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 data leakage, 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 Minimos Cuadrados Ordinarios (OLS) 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` $(.)$

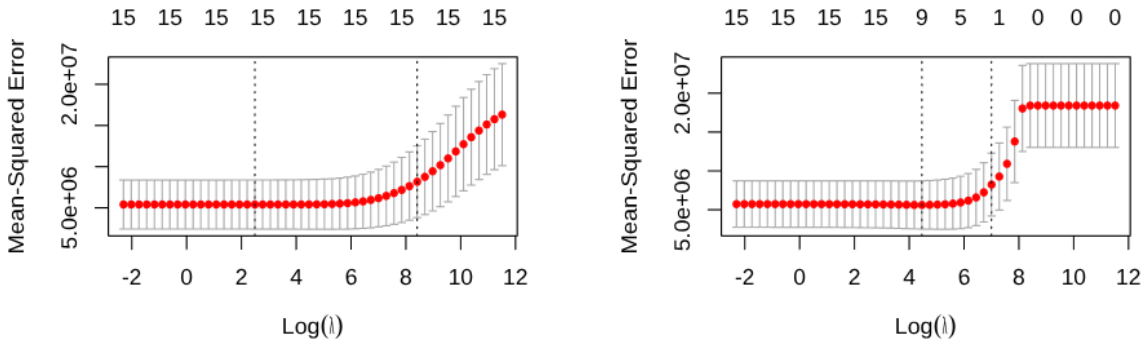
mientras que las demás variables no mostraron significancia estadística en el modelo ajustado. Al evaluar el rendimiento de este modelo en el conjunto de prueba, se obtuvo un Error Cuadrático Medio (MSE) de **2,551,734**. Esto corresponde a un Error Cuadrático Medio Raíz (RMSE) de **1,597.4**.

1.3 Inciso c)

Se ajustó una regresión *Ridge* con validación cruzada de 5 pliegues para seleccionar el parámetro de regularización λ. El valor de λ que minimiza el error $\lambda_{\text{mín}}$ elegido por CV fue **12.07**, y $\lambda_{1se} =$ **4498.43** como alternativa más parsimoniosa (ver Figura 1.1). 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, correspondientes a las variables: `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. Las líneas verticales indican $\lambda_{\text{mín}}$ y λ_{1se}.

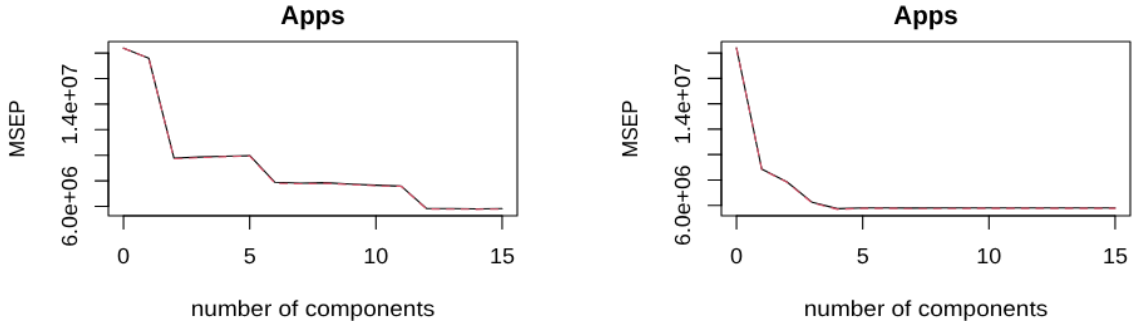
1.5 Inciso e)

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`. Como se espera, aumentar el número de componentes no mejora la explicación de `Apps` significativamente.

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.

En esta ocasión, el modelo con 4 componentes explica el 72.01 % de la varianza en `Apps`, lo que indica que PLSR logra una mejor explicación de la variable respuesta con menos componentes en comparación con PCR.

1.7 Inciso g)

Para entender el desempeño de los modelos, debemos entender la variable de respuesta `Apps`. Esta variable tiene un rango amplio, desde un mínimo de 81 hasta un máximo de 48,094 solicitudes, con un promedio de aproximadamente 3,002 y una mediana de 1,558. La distribución es altamente sesgada a la derecha, con algunas universidades recibiendo un número excepcionalmente alto de solicitudes.

Cuadro 1.1: Errores de prueba por modelo

Modelo	**MSE**	**RMSE**
OLS	2,551,734	1,597.4
Ridge ($\lambda_{\text{mín}} = 12.07$)	2,530,947	1,590.9
Lasso ($\lambda_{\text{mín}} = 86.85$)	2,395,665	1,547.8
PCR ($M = 12$)	**2,389,249**	**1,545.7**
PLSR ($M = 4$)	2,457,676	1,567.7

En términos de precisión, todos los métodos logran errores de prueba de magnitud similar: RMSE entre $\approx$ 1545 y 1600 solicitudes. Dado que el promedio de `Apps` ronda $\sim$ 3000, 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 (Cuadro 1.1), **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.

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 $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ 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.

Primero, analicemos las condiciones dadas:

- $n = 2, p = 2$.
- $x_{11} = x_{12}$ y $x_{21} = x_{22}$. Esto significa que la columna 1 (X_1) y la columna 2 (X_2) de la matriz X son idénticas: $X_1 = X_2$. Estamos en un caso de colinealidad perfecta.
- $y_1 + y_2 = 0$ y $x_{11} + x_{21} = 0$. Dado que $\hat{\gamma}_0 = 0$, esto implica que las variables (y, X_1, X_2) están centradas.
- Para simplificar, definamos $a = x_{11} = x_{12}$ y $c = y_1$.
- De las condiciones:
  - $x_{21} = -x_{11} = -a$
  - $x_{22} = -x_{12} = -a$ (consistente con $x_{21} = x_{22}$)
  - $y_2 = -y_1 = -c$

Nuestros datos son:

$$y = \begin{pmatrix} c \\ -c \end{pmatrix}, \quad X = \begin{pmatrix} a & a \\ -a & -a \end{pmatrix}$$

La suma de cuadrados residuales (RSS) es:

$$RSS(\beta_1, \beta_2) = \sum_{i=1}^{2} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2$$

$$RSS = (y_1 - (x_{11}\beta_1 + x_{12}\beta_2))^2 + (y_2 - (x_{21}\beta_1 + x_{22}\beta_2))^2$$

Sustituyendo nuestros valores:

$$\begin{aligned}
RSS &= (c - (a\beta_1 + a\beta_2))^2 + (-c - (-a\beta_1 - a\beta_2))^2 \\
RSS &= (c - a(\beta_1 + \beta_2))^2 + (-c + a(\beta_1 + \beta_2))^2 \\
RSS &= (c - a(\beta_1 + \beta_2))^2 + (-(c - a(\beta_1 + \beta_2)))^2 \\
RSS &= 2(c - a(\beta_1 + \beta_2))^2
\end{aligned}$$

Como podemos ver, el RSS solo depende de la suma de los coeficientes, $S = \beta_1 + \beta_2$.

2.1 Inciso a)

La regresión ridge busca minimizar el RSS más una penalización L2:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Bajo las suposiciones dadas y con $\hat{\gamma}_0 = 0$, el problema de optimización para β_1 y β_2 es:

$$\min_{\beta_1, \beta_2} \left\{ (y_1 - x_{11}\beta_1 - x_{12}\beta_2)^2 + (y_2 - x_{21}\beta_1 - x_{22}\beta_2)^2 + \lambda(\beta_1^2 + \beta_2^2) \right\}$$

Sustituyendo la forma simplificada del RSS que encontramos:

$$\min_{\beta_1, \beta_2} \left\{ 2(c - a(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2) \right\}$$

(Donde $c = y_1$ y $a = x_{11}$).

2.2 Inciso b)

El argumento se basa en la **simetría** de la función objetivo de ridge y la **unicidad** de la penalización L2.

1. **Simetría:** Sea la función objetivo $L_{Ridge}(\beta_1, \beta_2) = 2(c - a(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2)$.
   - El término RSS, $2(c - a(\beta_1 + \beta_2))^2$, es simétrico respecto a β_1 y β_2. Si los intercambiamos, el valor no cambia.
   - El término de penalización, $\lambda(\beta_1^2 + \beta_2^2)$, también es simétrico.
   - Por lo tanto, la función objetivo total $L_{Ridge}(\beta_1, \beta_2)$ es simétrica.
2. **Unicidad del Mínimo:** Para $\lambda > 0$, la función L_{Ridge} es estrictamente convexa. Esto se debe a que la penalización L2 (una esfera) es estrictamente convexa. Esto garantiza que existe un *único* mínimo global $(\hat{\beta}_1, \hat{\beta}_2)$.

Dado que la función es simétrica y tiene un único mínimo, el minimizador en sí debe ser simétrico, lo que implica $\hat{\beta}_1 = \hat{\beta}_2$.

2.3 Inciso c)

La regresión lasso busca minimizar el RSS más una penalización L1:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Bajo las suposiciones dadas:

$$\min_{\beta_1,\beta_2} \left\{ (y_1 - x_{11}\beta_1 - x_{12}\beta_2)^2 + (y_2 - x_{21}\beta_1 - x_{22}\beta_2)^2 + \lambda(|\beta_1| + |\beta_2|) \right\}$$

Sustituyendo la forma simplificada del RSS:

$$\min_{\beta_1,\beta_2} \left\{ 2(c - a(\beta_1 + \beta_2))^2 + \lambda(|\beta_1| + |\beta_2|) \right\}$$

2.4 Inciso d)

A diferencia de Ridge, la penalización L1 no produce un mínimo único en este escenario. El argumento es el siguiente:

1. **Descomposición del Problema:** Al igual que con Ridge, podemos descomponer el problema. Sea $S = \beta_1 + \beta_2$.

$$\min_{S} \left\{ 2(c - aS)^2 + \min_{\beta_1+\beta_2=S} (\lambda(|\beta_1| + |\beta_2|)) \right\}$$

2. **Problema Interno (Penalización L1):** Consideremos el problema interno mín$(|\beta_1| + |\beta_2|)$ sujeto a $\beta_1 + \beta_2 = S$.
   - El valor mínimo de esta penalización es $\lambda|S|$.
   - Sin embargo, este mínimo **no es único**. Se alcanza para *cualquier* par (β_1, β_2) tal que $\beta_1 + \beta_2 = S$ y $\beta_1 \cdot \beta_2 \geq 0$ (es decir, β_1 y β_2 tienen el mismo signo o uno es cero).
   - Por ejemplo, si $S = 5$, el mínimo de penalización (5) se alcanza en $(5, 0)$, $(0, 5)$, $(2, 3)$, $(4, 1)$, etc.

3. **Problema Externo (Encontrar $\hat{S}$):** El problema se reduce a encontrar la *suma* óptima $\hat{S}$ minimizando:

$$\min_{S} \left\{ 2(c - aS)^2 + \lambda|S| \right\}$$

   Este es un problema de optimización 1D (un "Lasso 1D") que tiene una solución **única** para la suma, $\hat{S}$. Esta solución $\hat{S}$ es el resultado de aplicar *soft-thresholding* a la solución OLS ($S_{OLS} = c/a$).

4. **Descripción de las Soluciones:**

   Una vez que se encuentra el valor único $\hat{S}$, la solución $(\hat{\beta}_1, \hat{\beta}_2)$ es *cualquier* par que cumpla:

   a) $\hat{\beta}_1 + \hat{\beta}_2 = \hat{S}$

   b) $|\hat{\beta}_1| + |\hat{\beta}_2| = |\hat{S}|$

   Esto describe un conjunto de soluciones (un cojunto convexo):

   - **Si** $\hat{S} > 0$**:** El conjunto de soluciones es el segmento de línea que conecta $(\hat{S}, 0)$ y $(0, \hat{S})$. Es decir, todos los puntos $(\beta_1, \hat{S} - \beta_1)$ para $\beta_1 \in [0, \hat{S}]$.
   - **Si** $\hat{S} < 0$**:** El conjunto de soluciones es el segmento de línea que conecta $(\hat{S}, 0)$ y $(0, \hat{S})$. Es decir, todos los puntos $(\beta_1, \hat{S} - \beta_1)$ para $\beta_1 \in [\hat{S}, 0]$.
   - **Si** $\hat{S} = 0$**:** (Es decir, si λ es suficientemente grande), la solución es única: $(\hat{\beta}_1, \hat{\beta}_2) = (0, 0)$.

En resumen, mientras que Ridge (L2) identifica la solución simétrica $\hat{\beta}_1 = \hat{\beta}_2$ como la única óptima para una suma dada, Lasso (L1) considera que *todas* las soluciones en el mismo signo (incluyendo las soluciones *sparse* como $(\hat{S}, 0)$ o $(0, \hat{S})$) son igualmente óptimas.