Laboratorio 1: Regresión Lineal y Regularización (L2 y L1)

Duración: 2 horas Formato: Competencia por equipos

Objetivo: Que los estudiantes implementen desde cero un modelo de regresión lineal con regularización L2 (Ridge) y L1 (Lasso), analicen los efectos del ajuste de hiperparámetros y visualicen los fenómenos de sesgo, varianza, selección de variables y estabilidad numérica.

Estructura detallada de la práctica

Tiempo	Actividad
0:00–0:10	Bienvenida + Formación de equipos (3–4 personas).
0:10–0:30	Fase 1: Implementación base de Regresión Lineal por mínimos cuadrados.
0:30–0:45	Fase 2: Regularización L2 (Ridge) y análisis de estabilidad.
0:45–1:00	Fase 3: Regularización L1 (Lasso) y selección de variables.
1:00–1:30	Fase 4: Desafío por equipos: *"¡Predice mejor con menos!"*.
1:30–1:50	Fase 5: Presentaciones rápidas por equipo (3 minutos c/u).
1:50–2:00	Cierre, feedback y selección del "modelo ganador".

Fase 1 – Implementación base

Objetivo: Implementar el estimador de mínimos cuadrados:

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Dataset inicial: Versión extendida del ejercicio de clase (*Demanda de energía*) con más ejemplos, ruido y colinealidad artificial.

Validaciones a observar:

- MSE en entrenamiento y validación.
- Estabilidad del sistema (condición de la matriz).
- Cuidado con matrices no invertibles.

Fase 2 – Regularización L2 (Ridge)

Objetivo: Implementar el estimador Ridge:

$$\hat{\theta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Exploración:

- Cómo cambia la magnitud de los coeficientes con λ.
- Comparación visual del ajuste.
- ¿Se reduce el sobreajuste? ¿Se mejora la estabilidad?

Fase 3 – Regularización L1 (Lasso)

Objetivo: Resolver:

$$\hat{\theta}_{lasso} = \arg\min_{\theta} \left(\frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right)$$

Hint: La actualización de cada coeficiente se realiza mediante la regla de **soft-thresholding**, la cual puede expresarse de manera sencilla como:

$$S(z, \gamma) = \begin{cases} z - \gamma & \text{si } z > \gamma \\ z + \gamma & \text{si } z < -\gamma \\ 0 & \text{si } |z| \leq \gamma \end{cases}$$

donde z es el valor intermedio calculado para el coeficiente y γ es el **umbral de corte**, proporcional al parámetro de regularización λ y a la escala de los datos. En la implementación clásica de *coordinate descent* para Lasso se utiliza:

$$\gamma = \frac{\lambda}{2n}$$

donde n es el número de observaciones.

Exploración:

- ¿Cuántos coeficientes se van a cero?
- ¿Qué tan agresiva es la selección?

Fase 4 – Desafío: "¡Predice mejor con menos!"

Dataset oculto: Similar al anterior, pero con más ruido y variables redundantes.

Reglas:

- Evaluación con MSE en test + penalización por número de coeficientes distintos de cero.
- Puntos extra si el modelo es interpretable o logra buena generalización.
- Tiempo límite: 25–30 minutos.

Fase 5 – Presentaciones rápidas

Cada equipo tendrá 3 minutos para explicar:

- Qué modelo usaron y por qué.
- Qué observaron al ajustar λ.
- Qué harían distinto.

Criterios de selección del modelo ganador:

- Menor MSE con buena justificación.
- Mejor uso de la regularización.
- Creatividad en la visualización o análisis.

Bonus – Elastic Net (Regularización L1 + L2)

Como reto adicional, los equipos pueden implementar el modelo **Elastic Net**, que combina las penalizaciones L1 (Lasso) y L2 (Ridge) en una sola formulación:

$$\hat{\theta}_{EN} = \arg\min_{\theta} \left(\frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \Big(\alpha \|\theta\|_1 + \frac{1-\alpha}{2} \|\theta\|_2^2 \Big) \right)$$

donde:

- λ controla la fuerza total de regularización.
- $\alpha \in [0, 1]$ equilibra la mezcla entre L1 y L2:
  - $\alpha = 1$: Lasso puro.
  - $\alpha = 0$: Ridge puro.
  - $0 < \alpha < 1$: combinación.

Idea principal:

- L1 selecciona variables (lleva coeficientes a cero).
- L2 estabiliza en presencia de multicolinealidad y evita coeficientes extremos.
- Elastic Net ofrece lo mejor de ambos mundos.

Incentivo: Los equipos que entreguen una implementación funcional de **Elastic Net** obtendrán un **bonus competitivo del 15%** aplicado sobre su *score final* en el tablero. Este bono no altera la calificación académica oficial, pero sí les permitirá escalar puestos adicionales en el ranking de la competencia.

Rúbrica de Evaluación de la Práctica

La dinámica del laboratorio se desarrolla en fases con tiempos definidos. Se da el banderazo de cada fase para guiar el ritmo de trabajo, pero la calificación se otorga únicamente al final, con base en la siguiente rúbrica:

Criterio	Ponderación
Implementación funcional de modelos (OLS, Ridge y Lasso)	40%
Calidad del análisis y visualización (gráficas, interpretación de λ, claridad de resultados)	20%
Capacidad de generalización en el conjunto oculto	20%
Claridad y creatividad en la presentación final (síntesis, argumentación, conclusiones)	20%
Total	**100%**

Condición importante: La práctica **solo será calificada** al equipo que entregue su cuaderno completo (`.ipynb`) con el desarrollo de la práctica, incluyendo en la primera sección los **nombres de todos los integrantes del equipo**.

Formato de Entrega de Resultados de la Competencia

Cada equipo deberá entregar un archivo `.csv` con las predicciones para el conjunto oculto `hidden_test.csv`.

Requisitos mínimos:

- Una columna obligatoria: `y_pred`, con tantas filas como ejemplos tenga el conjunto oculto.

Campos opcionales (recomendados):

- `team`: nombre del equipo (solo en la primera fila).
- `model`: breve descripción del modelo usado.
- `nnz`: número de coeficientes distintos de cero (para aplicar penalización opcional).

Ejemplo de formato válido:

```
y_pred,team,model,nnz
121.7,Equipo_Beta,"Lasso lambda=0.1",5
119.9,,,
129.3,,,
...
```

El ranking final se calculará con base en el **MSE en el conjunto oculto**, y en caso de aplicar penalizaciones, se sumará un término proporcional a `nnz`.

Nota importante: Es **obligatorio entregar los resultados del modelo en formato .csv**. En caso de no hacerlo, se aplicará una **penalización del 5%** sobre la calificación final de la práctica.