Métodos de Regularización en Aprendizaje Automático

Informe Técnico
FERNANDO JOSE MAMANI MACHACA

19 de febrero de 2025

1. Introducción

La regularización es una técnica fundamental en el aprendizaje automático que ayuda a prevenir el sobreajuste (overfitting) en modelos estadísticos. Este informe presenta los principales métodos de regularización utilizados en la actualidad.

2. Regularización L1 (Lasso)

La regularización L1, también conocida como Lasso (Least Absolute Shrinkage and Selection Operator), añade el valor absoluto de los coeficientes como penalización a la función de pérdida:

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

Donde:

- λ es el parámetro de regularización
- β_j son los coeficientes del modelo
- x_i son las variables de entrada
- y_i son las variables objetivo

3. Regularización L2 (Ridge)

La regularización L2 o Ridge añade el cuadrado de la magnitud de los coeficientes como término de penalización:

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

4. Ejemplo

Prueba del método L2 (Ridge) con un conjunto de datos (dataset):

Región	Acelga	Aji	Ajo	Albahaca	Apio	Arracacha	Arroz cáscara	Arveja grano	Arveja grano verde	Avena grano
Nacional	**5,932**	**48,337**	**113,732**	**5,912**	**36,061**	**17,403**	**3,439,013**	**51,278**	**148,940**	**23,840**
Amazonas	40	207	18	0	0	9,500	400,726	100	2,042	0
Áncash	639	1,647	377	0	758	0	60,445	494	3,641	0
Apurímac	0	63	470	74	23	0	0	2,705	4,363	1,616
Arequipa	1,652	1,408	78,619	1,764	9,956	0	280,697	5	16,248	0
Ayacucho	0	65	2,496	0	1,052	0	12	6,171	6,049	641
Cajamarca	0	169	7,219	0	0	7,444	186,783	14,041	26,202	0
Callao	0	0	0	0	0	0	0	0	0	0
Cusco	0	0	0	0	0	0	1,985	2,897	3,041	9,500
Huancavelica	0	0	1,064	0	0	0	0	5,491	22,550	866
Huánuco	0	620	243	0	205	0	56,046	1,618	14,290	2,011
Ica	39	5,817	0	313	0	0	0	0	352	0
Junín	1,175	1,313	4,841	0	2,265	0	1,029	1,405	33,263	680
La Libertad	229	4,120	1,469	0	2,175	0	309,349	10,839	5,132	0
Lambayeque	0	0	0	0	0	0	440,156	413	3,925	0
Lima	0	8,298	15,835	0	7,369	0	0	35	3,103	0
Lima Metropolitana	2,126	4,845	112	3,761	9,646	0	0	0	235	0
Loreto	0	741	0	0	0	0	108,073	0	0	0
Madre de Dios	0	0	0	0	0	0	7,608	0	0	0
Moquegua	0	0	262	0	0	68	0	0	197	0
Pasco	0	4,874	10	0	0	0	2,036	103	2,700	0
Piura	0	0	283	0	0	0	482,421	3,908	1,458	0
Puno	0	0	180	0	0	390	89	1,051	0	8,527
San Martín	0	381	0	0	0	0	868,384	0	0	0
Tacna	32	13,574	235	0	2,612	0	0	0	149	0
Tumbes	0	11	0	0	0	0	128,633	0	0	0
Ucayali	0	184	0	0	0	0	104,542	0	0	0

(sigue)

Figura 1: Dataset inicial para análisis con método Ridge.

4.1. Resultados del Método después de ejecutarlo

Escojer uno de los productos disponibles a analizar

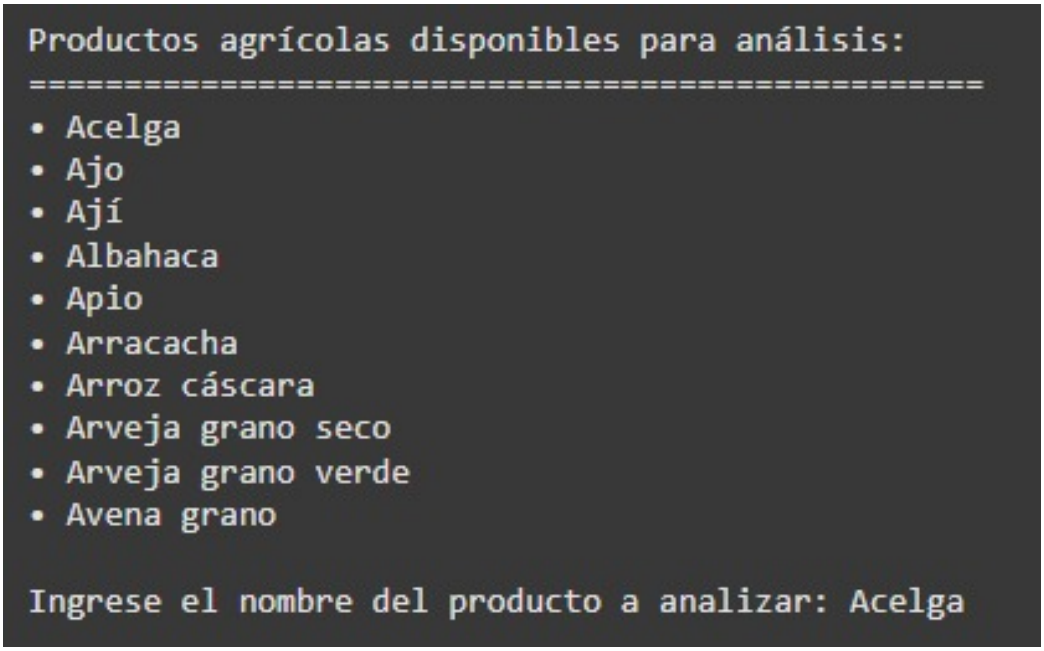

Figura 2: Análisis preliminar del dataset.

4.2. Resultados después del análisis

Resultados analizados después de escoger uno de los productos disponibles.

```
Iniciando análisis...

Resultados del Análisis - Acelga
==================================================
Mejor valor de alpha: 100.0
Error cuadrático medio (MSE): 3914986271.12
Coeficiente de determinación (R²): 0.2958

Influencia de cada región en la producción:
==================================================
Región_Nacional               20774.384705
Región_Junín                   7251.366116
Región_Cajamarca                988.357235
Región_San Martín               508.278599
Región_Lima                   -1011.970383
Región_Lambayeque             -1281.398961
Región_Puno                   -1480.428496
Región_Cusco                  -1694.611763
Región_La Libertad            -1705.848055
Región_Ica                    -1718.926537
Región_Huánuco                -1731.910273
Región_Madre de Dios          -1792.571804
Región_Ucayali                -1793.699847
Región_Tacna                  -1821.400329
Región_Arequipa               -1835.591557
Región_Áncash                 -1901.551525
Región_Apurímac               -2001.200806
Región_Lima Metropolitana     -2014.120112
Región_Pasco                  -2118.510369
Región_Piura                  -2120.974011
Región_Huancavelica           -2124.458542
Región_Loreto                 -2174.190600
Región_Ayacucho               -2178.427812
Región_Tumbes                 -2288.209704
Región_Moquegua               -2302.902159
Región_Callao                 -2314.370070
```

Figura 3: Resultados finales del análisis Ridge.

4.3. Interpetacion de los resultados

El modelo no es muy preciso para predecir la producción de acelga, como lo indica el bajo R^2 y el alto MSE.

La Región Nacional y Junín son las que más contribuyen a la producción, mientras que otras regiones tienen un impacto negativo o insignificante.

Se recomienda explorar más variables (como clima, tipo de suelo, técnicas de cultivo) para mejorar la precisión del modelo.

4.4. Enlace al Código en Google Colab

https://colab.research.google.com/drive/1p$_x$*CuhjnhwjMtRlHz7hVJqefn*4*clUQ*6*s*?*usp* = *sharing*

5. Regularización Elástica Net

Elastic Net combina las regularizaciones L1 y L2:

$$L(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2$$

6. Dropout

El Dropout es una técnica de regularización específica para redes neuronales que consiste en desactivar aleatoriamente un porcentaje de neuronas durante el entrenamiento. La probabilidad de mantener una neurona activa se denota como p, típicamente entre 0.5 y 0.8.

7. Early Stopping

Early Stopping es una técnica que detiene el entrenamiento cuando el rendimiento en el conjunto de validación comienza a deteriorarse, evitando así el sobreajuste.

8. Ventajas y Desventajas

8.1. Regularización L1

Ventajas:

- Produce modelos dispersos (sparse)
- Realiza selección de características

Desventajas:

- Puede ser inestable con características altamente correlacionadas
- No tiene solución analítica

8.2. Regularización L2

Ventajas:

- Tiene solución analítica
- Maneja bien características correlacionadas

Desventajas:

- No produce modelos dispersos
- No realiza selección de características

9. Conclusiones

La elección del método de regularización depende del problema específico:

- L1 es preferible cuando se busca sparsity y selección de características
- L2 es mejor para tratar con características correlacionadas
- Elastic Net proporciona un buen compromiso entre L1 y L2
- Dropout es específico para redes neuronales profundas
- Early Stopping es una técnica general aplicable a diversos algoritmos