ANÁLISIS DE REGRESIÓN LINEAL

Dataset Iris Flower

ITESO, Universidad Jesuita de Guadalajara

Laboratorio Aprendizaje Estadístico

Integrantes:

Francisco Tinoco

Juan Pedro Ley

David Rangel

Febrero 2026

Contenido

1. Objetivos 3

1.1 Objetivo General 3

1.2 Objetivos Específicos 3

2. Marco Teórico 4

2.1 Regresión Lineal 4

2.2 Regresión Polinomial 5

2.3 Interacción de Factores 5

2.4 Significancia de Factores 6

2.5 Regularización (Ridge, Lasso, ElasticNet) 7

2.6 Dataset Iris 8

3. Análisis del Dataset 9

4. Metodología y Pipeline 10

5. Modelos Propuestos 11

6. Resultados y Comparaciones 12

7. Análisis de Significancia 14

8. Conclusiones 15

9. Referencias 16

1. Objetivos

1.1 Objetivo General

Realizar un análisis completo de regresión lineal sobre el dataset Iris para identificar las relaciones entre las características morfológicas de las flores y evaluar la capacidad predictiva de diferentes modelos de regresión, incluyendo técnicas de regularización.

1.2 Objetivos Específicos

- Aplicar técnicas de preprocesamiento de datos incluyendo transformación de variables categóricas mediante One-Hot Encoding y escalamiento de variables numéricas.
- Desarrollar tres modelos de regresión para predecir diferentes características de las flores: longitud del pétalo, ancho del pétalo y longitud del sépalo.
- Implementar y comparar cuatro variantes de regresión: OLS (sin penalización), Ridge (L2), Lasso (L1) y ElasticNet.
- Evaluar el rendimiento de los modelos mediante el coeficiente de determinación R^2 tanto en datos de entrenamiento como de prueba.
- Realizar un análisis de significancia estadística para identificar qué factores tienen un impacto significativo en cada variable de respuesta.
- Interpretar los resultados obtenidos y formular conclusiones fundamentadas sobre las relaciones entre las variables del dataset.

2. Marco Teórico

2.1 Regresión Lineal

La regresión lineal es una técnica estadística que permite modelar la relación entre una variable dependiente (Y) y una o más variables independientes (X). El objetivo es encontrar la línea (o hiperplano en el caso multivariado) que mejor se ajuste a los datos observados.

Modelo de Regresión Lineal Simple:

$$\hat{y} = \beta_0 + \beta_1 x$$

Modelo de Regresión Lineal Múltiple:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Donde:

- $\hat{y}$ es el valor predicho de la variable dependiente
- β_0 es el intercepto (valor de Y cuando todas las X son cero)
- β_j son los coeficientes de regresión (efecto de cada variable X_j)
- x_j son las variables independientes

Los coeficientes se estiman mediante el método de Mínimos Cuadrados Ordinarios (OLS), que minimiza la suma de los errores al cuadrado (RSS):

$$RSS = \Sigma(y_i - \hat{y}_i)^2$$

2.2 Regresión Polinomial

La regresión polinomial es una extensión de la regresión lineal que permite modelar relaciones no lineales entre las variables. Se logra incluyendo potencias de las variables independientes como predictores adicionales.

Modelo Polinomial de Grado d:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_a x^d$$

Consideraciones importantes:

- A mayor grado del polinomio, mayor flexibilidad del modelo pero también mayor riesgo de sobreajuste (overfitting).
- Es fundamental evaluar el modelo en datos de prueba para detectar sobreajuste.
- La elección del grado óptimo puede realizarse mediante validación cruzada.

2.3 Interacción de Factores

La interacción de factores ocurre cuando el efecto de una variable independiente sobre la variable dependiente depende del valor de otra variable independiente. En términos del modelo, esto se representa multiplicando las variables:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2)$$

El término β_3 captura el efecto de interacción. Si β_3 es estadísticamente significativo, indica que el efecto de x_1 varía según el nivel de x_2 (y viceversa).

2.4 Significancia de Factores

El análisis de significancia permite determinar si los coeficientes estimados son estadísticamente diferentes de cero, es decir, si las variables independientes tienen un efecto real sobre la variable dependiente.

Prueba de Hipótesis:

H_0: $\beta_j = 0$ (no hay efecto)

H_1: $\beta_j \neq 0$ (sí hay efecto)

Estadístico t:

$$t = \hat{\beta}_j / SE(\hat{\beta}_j)$$

Donde $SE(\hat{\beta}_j)$ es el error estándar del coeficiente estimado.

Criterio de decisión:

- Si p-value < 0.05, se rechaza H_0 y el factor es estadísticamente significativo.
- Si p-value ≥ 0.05, no se rechaza H_0 y el factor no es estadísticamente significativo.
- El intervalo de confianza del 95% se calcula como: $\hat{\beta}_j \pm 2 \cdot SE(\hat{\beta}_j)$

2.5 Regularización (Ridge, Lasso, ElasticNet)

La regularización es una técnica que añade una penalización a la función de costo para controlar la magnitud de los coeficientes y prevenir el sobreajuste.

Ridge (L2):

$$Costo = RSS + \lambda \cdot \Sigma \beta_j^2$$

La penalización L2 reduce los coeficientes pero nunca los hace exactamente cero. Es útil cuando todas las variables son potencialmente relevantes.

Lasso (L1):

$$\text{Costo} = \text{RSS} + \lambda \cdot \Sigma|\beta_j|$$

La penalización L1 puede hacer que algunos coeficientes sean exactamente cero, realizando así selección automática de variables.

ElasticNet:

$$\text{Costo} = \text{RSS} + \lambda_1 \cdot \Sigma|\beta_j| + \lambda_2 \cdot \Sigma\beta_j^2$$

Combina las penalizaciones L1 y L2, obteniendo las ventajas de ambos métodos.

Hiperparámetro λ (alpha):

- λ = 0: equivale a regresión OLS sin penalización
- λ pequeño: penalización leve, coeficientes cercanos a OLS
- λ grande: penalización fuerte, coeficientes cercanos a cero

2.6 Dataset Iris

El dataset Iris es uno de los conjuntos de datos más conocidos en el campo del aprendizaje automático y la estadística. Fue introducido por el estadístico y biólogo británico Ronald Fisher en 1936 en su paper "The use of multiple measurements in taxonomic problems".

Características del dataset:

- 150 muestras en total
- 3 especies de Iris: Setosa, Versicolor y Virginica
- 50 muestras por especie
- 4 características numéricas medidas en centímetros

Variables del dataset:

Variable	Tipo	Descripción
sepal_length	Numérica	Longitud del sépalo (cm)
sepal_width	Numérica	Ancho del sépalo (cm)
petal_length	Numérica	Longitud del pétalo (cm)
petal_width	Numérica	Ancho del pétalo (cm)
species	Categórica	Especie de la flor

3. Análisis del Dataset

3.1 Origen de los Datos

Los datos fueron obtenidos de Kaggle, específicamente del repositorio 'Iris Flower Dataset'. Originalmente, los datos provienen del UCI Machine Learning Repository.

3.2 Estadísticas Descriptivas

Variable	Min	Media	Max	Std
sepal_length	4.3	5.84	7.9	0.83
sepal_width	2.0	3.05	4.4	0.43
petal_length	1.0	3.76	6.9	1.76
petal_width	0.1	1.20	2.5	0.76

3.3 Transformaciones Necesarias

One-Hot Encoding para species: La variable categórica 'species' se transforma en variables dummy binarias para poder incluirla en el modelo de regresión.

Escalamiento StandardScaler: Se aplica normalización (media=0, std=1) a las variables numéricas, especialmente importante para los modelos con regularización.

4. Metodología y Pipeline

El proceso de análisis sigue un pipeline estructurado que garantiza la reproducibilidad y validez de los resultados:

Paso 1 - Carga de Datos: Lectura del archivo CSV con pandas.

Paso 2 - Limpieza: Verificación de valores nulos y tipos de datos.

Paso 3 - Transformación: One-Hot Encoding para variables categóricas.

Paso 4 - Separación de Variables: Definición de X (predictores) e Y (variable respuesta).

Paso 5 - Train-Test Split: División 70% entrenamiento, 30% prueba.

Paso 6 - Escalamiento: Aplicación de StandardScaler a los datos.

Paso 7 - Ajuste de Modelos: Entrenamiento de OLS, Ridge, Lasso y ElasticNet.

Paso 8 - Evaluación: Cálculo de R^2 en train y test.

Paso 9 - Análisis de Significancia: Obtención de p-values con statsmodels.

Paso 10 - Conclusiones: Interpretación de resultados.

5. Modelos Propuestos

Se proponen tres modelos de regresión, cada uno con cuatro versiones (OLS, Ridge, Lasso, ElasticNet):

5.1 Modelo 1: Predecir petal_length

$$\text{petal\_length} = \beta_0 + \beta_1(\text{sepal\_length}) + \beta_2(\text{sepal\_width}) + \beta_3(\text{petal\_width}) + \beta_4(\text{versicolor}) + \beta_5(\text{virginica})$$

5.2 Modelo 2: Predecir petal_width

$$\text{petal\_width} = \beta_0 + \beta_1(\text{sepal\_length}) + \beta_2(\text{sepal\_width}) + \beta_3(\text{petal\_length}) + \beta_4(\text{versicolor}) + \beta_5(\text{virginica})$$

5.3 Modelo 3: Predecir sepal_length

$$\text{sepal\_length} = \beta_0 + \beta_1(\text{sepal\_width}) + \beta_2(\text{petal\_length}) + \beta_3(\text{petal\_width}) + \beta_4(\text{versicolor}) + \beta_5(\text{virginica})$$

5.4 Versiones de Cada Modelo

Versión	Tipo	Función de Costo
1	OLS (Sin penalización)	RSS
2	Ridge (L2)	$RSS + \lambda\Sigma\beta_j^2$
3	Lasso (L1)	$RSS + \lambda\Sigma\lvert\beta_j\rvert$
4	ElasticNet	$RSS + \lambda_1\Sigma\lvert\beta_j\rvert + \lambda_2\Sigma\beta_j^2$

6. Resultados y Comparaciones

6.1 Resultados del Modelo 1 (petal_length)

Tipo	R^2 Train	R^2 Test
OLS	~0.98	~0.96
Ridge	~0.98	~0.96
Lasso	~0.97	~0.95
ElasticNet	~0.97	~0.95

6.2 Resultados del Modelo 2 (petal_width)

Tipo	R^2 Train	R^2 Test
OLS	~0.95	~0.94

Ridge	~0.95	~0.94
Lasso	~0.94	~0.93
ElasticNet	~0.94	~0.93

6.3 Resultados del Modelo 3 (sepal_length)

Tipo	R^2 Train	R^2 Test
OLS	~0.85	~0.78
Ridge	~0.85	~0.78
Lasso	~0.84	~0.77
ElasticNet	~0.84	~0.77

6.4 Análisis Comparativo

Observaciones principales:

- Los modelos 1 y 2 (predicción de pétalos) obtienen R^2 muy altos (>0.93), indicando que las dimensiones de los pétalos están altamente correlacionadas.
- El modelo 3 (predicción de sepal_length) tiene R^2 más bajo (~0.78), sugiriendo menor correlación con las demás variables.
- Los modelos penalizados (Ridge, Lasso, ElasticNet) no mejoran significativamente sobre OLS, indicando ausencia de sobreajuste severo.
- La diferencia entre R^2 de entrenamiento y prueba es pequeña, confirmando buena generalización.

7. Análisis de Significancia

Se utilizó statsmodels para obtener los p-values de cada coeficiente en los modelos OLS. Un factor se considera estadísticamente significativo si su p-value es menor a 0.05.

7.1 Significancia en Modelo 1 (petal_length)

Variable	p-value	Significativo
sepal_length	> 0.05	No
sepal_width	> 0.05	No
petal_width	< 0.001	Sí ✓
species_versicolor	< 0.001	Sí ✓
species_virginica	< 0.001	Sí ✓

7.2 Significancia en Modelo 2 (petal_width)

Variable	p-value	Significativo
sepal_length	> 0.05	No

sepal_width	> 0.05	No
petal_length	< 0.001	Sí ✓
species_versicolor	> 0.05	No
species_virginica	< 0.05	Sí ✓

7.3 Significancia en Modelo 3 (sepal_length)

Variable	**p-value**	**Significativo**
sepal_width	< 0.001	Sí ✓
petal_length	< 0.001	Sí ✓
petal_width	> 0.05	No
species_versicolor	> 0.05	No
species_virginica	> 0.05	No

8. Conclusiones

A partir del análisis realizado sobre el dataset Iris, se obtienen las siguientes conclusiones:

8.1 Respecto al Objetivo General

Se logró realizar un análisis completo de regresión lineal sobre el dataset Iris, identificando las relaciones entre las características morfológicas de las flores. Los modelos desarrollados demuestran capacidades predictivas significativas, especialmente para las variables relacionadas con los pétalos.

8.2 Respecto a los Objetivos Específicos

Preprocesamiento: Se aplicó exitosamente One-Hot Encoding para la variable species y StandardScaler para las variables numéricas, permitiendo el correcto funcionamiento de todos los modelos.

Desarrollo de Modelos: Se implementaron tres modelos de regresión con cuatro variantes cada uno (OLS, Ridge, Lasso, ElasticNet), cumpliendo con la estructura propuesta.

Evaluación de Rendimiento: Los modelos de predicción de pétalos alcanzaron R^2 superiores a 0.93, mientras que la predicción de sepal_length obtuvo R^2 de aproximadamente 0.78.

Significancia Estadística: Se identificó que petal_width y las variables de especie son los factores más significativos para predecir petal_length, mientras que petal_length es el factor dominante para predecir petal_width.

Regularización: Los modelos penalizados no mostraron mejoras significativas sobre OLS, indicando que el dataset Iris, al ser compacto (150 muestras, 5 variables), no presenta problemas de sobreajuste.

8.3 Hallazgos Principales

- Las dimensiones de los pétalos (largo y ancho) están altamente correlacionadas entre sí, lo que permite predicciones muy precisas.
- La especie de la flor es un factor significativo para predecir características de los pétalos, reflejando las diferencias morfológicas entre las tres especies de Iris.
- Las características del sépalo tienen menor correlación con las demás variables, haciendo que sepal_length sea más difícil de predecir con precisión.

9. Referencias

- Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

- Kaggle. (s.f.). Iris Flower Dataset. Recuperado de https://www.kaggle.com/datasets/arshid/iris-flower-dataset

- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference, 57-61.