Liner and Non-Linear Regression

Angello Soldi* ANGELLO.SOLDI@UTEC.EDU.PE
UTEC, Peru

Fabian Castro* FABIAN.CASTRO@UTEC.EDU.PE
UTEC, Peru

John Monroy JOHN.MONROY@UTEC.EDU.PE
UTEC, Peru

Ingrid Santander * INGRID.SANTANDER@UTEC.EDU.PE
UTEC, Peru

Abstract

Este estudio analiza una serie temporal de mediciones climáticas aplicando regresión lineal y no lineal, con y sin regularización L1 (Lasso) y L2 (Ridge). El rendimiento de los modelos se evalúa mediante el error cuadrático medio (MSE) para identificar el enfoque más preciso. Los resultados comparan la efectividad de cada técnica en la predicción de datos climáticos.

Keywords: Regression, Regularization, MSE, Machine Learning

1. Introduction

El presente trabajo tiene como objetivo analizar una serie temporal unidimensional de mediciones climáticas a lo largo del tiempo mediante la implementación de técnicas de regresión, evaluando tanto modelos lineales como no lineales. Se busca identificar cuál de estos enfoques es más adecuado para capturar las tendencias inherentes a los datos y proporcionar predicciones precisas sobre las mediciones futuras. Para ello, se realizarán experimentos que involucren la comparación de modelos sin regularización y con regularización L1 (Lasso) y L2 (Ridge), con el propósito de analizar el impacto de cada método en la reducción de la complejidad del modelo, mitigando el sobreajuste y mejorando su capacidad de generalización. Como métrica de evaluación de la efectividad de los modelos se empleará el error cuadrático medio (MSE, por sus siglas en inglés), el cual permitirá medir la precisión de las predicciones en relación con los valores reales. Los resultados obtenidos serán analizados para establecer comparaciones entre los distintos enfoques y así determinar cuál ofrece un mejor rendimiento en el contexto específico de la serie temporal climática utilizada.

2. Data Exploration

Primero, se realizó un análisis visual de los datos de la serie temporal de mediciones climáticas mediante la creación de un gráfico de líneas. El eje horizontal (X) representa los días y el eje vertical (Y) muestra las temperaturas diarias registradas. En el gráfico se observa un patrón cíclico, lo cual sugiere una variación periódica en las temperaturas a lo largo del tiempo. Este tipo de comportamiento puede estar relacionado con cambios estacionales o fluctuaciones diarias en las condiciones climáticas. Las temperaturas oscilan entre aproximadamente 20°C y 80°C, lo que indica una amplitud significativa en las mediciones.

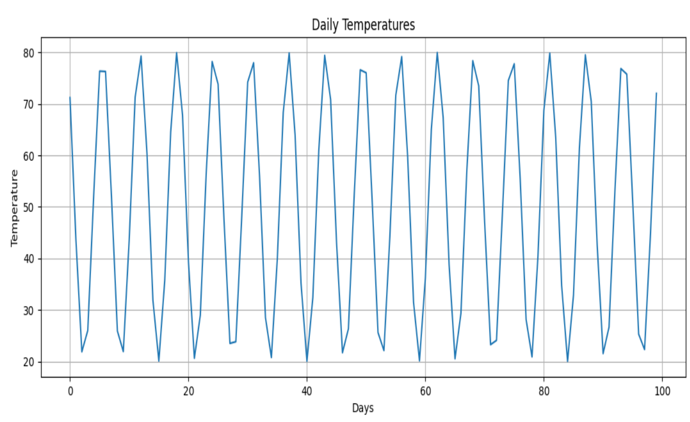

Figure 1: Gráfico de temperaturas diarias.

* These authors contributed equally

El análisis inicial del gráfico revela una tendencia a repetir ciclos regulares, lo que sugiere que un modelo de regresión no lineal podría ser más adecuado para capturar este tipo de comportamiento periódico. Sin embargo, también es posible explorar un modelo de regresión lineal como referencia para evaluar su capacidad de ajuste frente a esta serie temporal.

2.1. Basic Statistics

Statistic	Value
Count	100
Mean	49.835852
Standard Deviation	21.255238
Min	20.021208
25% Percentile	28.415242
50% Percentile	49.867323
75% Percentile	71.236622
Max	79.967635

Table 1: Estadísticas descriptivas de los datos de temperatura.

1. **Count (100):** Esto indica el número total de observaciones de temperatura en el conjunto de datos, que en este caso es 100.
2. **Media (49.84):** La media representa el valor promedio de la temperatura en todas las observaciones, siendo aproximadamente 49.84°C.
3. **Desviación Estándar (21.25):** La desviación estándar cuantifica la dispersión de los datos respecto a la media. Un valor de 21.25°C indica una variabilidad considerable en las temperaturas.
4. **Mínimo (20.02):** La temperatura mínima registrada es de 20.02°C, representando el valor más bajo observado en el conjunto de datos.
5. **Percentil 25% (28.41):** Este valor indica que el 25% de las temperaturas son inferiores a 28.41°C, lo que brinda información sobre el rango inferior de la distribución de los datos.
6. **Percentil 50% (Mediana, 49.87):** La mediana, de 49.87°C, representa el punto medio del conjunto de datos, donde la mitad de las temperaturas son menores y la otra mitad mayores. Es cercana a la media, lo que sugiere una distribución relativamente simétrica.
7. **Percentil 75% (71.24):** El percentil 75% indica que el 75% de los datos está por debajo de 71.24°C, lo que proporciona una idea del rango superior de la distribución de las temperaturas.
8. **Máximo (79.97):** La temperatura máxima registrada es de 79.97°C, siendo el valor más alto en el conjunto de datos.

2.2. Rolling Mean and Standard Deviation

El siguiente gráfico muestra las estadísticas de media móvil y desviación estándar móvil de las mediciones de temperatura. A continuación se describen las interpretaciones de cada una de las líneas:

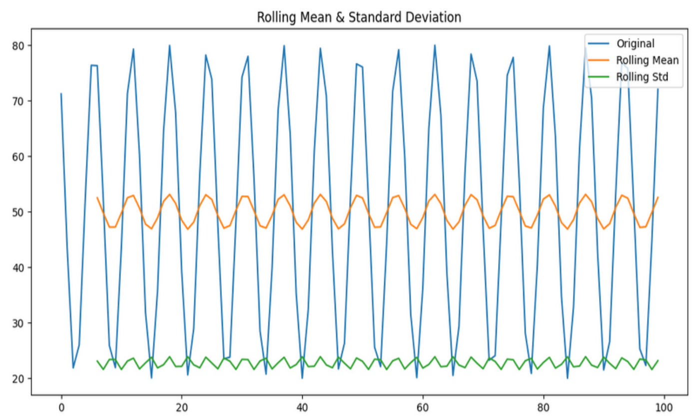

Figure 2: Gráfico de la Media Móvil y Desviación Estándar Móvil de la Temperatura.

1. **Serie Original (línea azul):** La línea azul representa la serie temporal original de las temperaturas. Se observa un patrón cíclico claramente definido, con fluctuaciones significativas entre 20°C y 80°C a intervalos regulares. Esto indica que los datos tienen una componente periódica, probablemente influenciada por ciclos estacionales o diarios.
2. **Media Móvil (línea naranja):** La media móvil suaviza las fluctuaciones de corto plazo en los datos y muestra la tendencia general del comportamiento de la temperatura a lo largo del tiempo. La media móvil sigue una trayectoria regular oscilando alrededor de los 50°C, lo cual

sugiere que las fluctuaciones significativas en la serie temporal se mantienen dentro de un rango estable, sin una tendencia ascendente o descendente clara a largo plazo.

3. **Desviación Estándar Móvil (línea verde):** La desviación estándar móvil refleja la volatilidad de las mediciones de temperatura en un intervalo de tiempo determinado (en este caso, 7 días). La línea verde permanece bastante baja y estable, lo que indica que la variabilidad a corto plazo dentro de cada ventana de 7 días es relativamente pequeña en comparación con la variabilidad observada en la serie original. Esto refuerza la idea de que la serie es cíclica y no muestra cambios significativos en la variación a lo largo del tiempo.

Conclusiones: El comportamiento cíclico de los datos se confirma tanto en la serie original como en la media móvil. La ausencia de grandes cambios en la desviación estándar móvil sugiere que el ciclo periódico es bastante regular, con poca variabilidad entre ciclos. Este análisis es útil para predecir futuras mediciones y ajustar modelos de regresión, especialmente si buscamos capturar estos patrones cíclicos.

2.3. Important Graphs

2.3.1. Histogram

El siguiente gráfico muestra un histograma de las temperaturas con una curva de estimación de densidad (KDE) superpuesta.

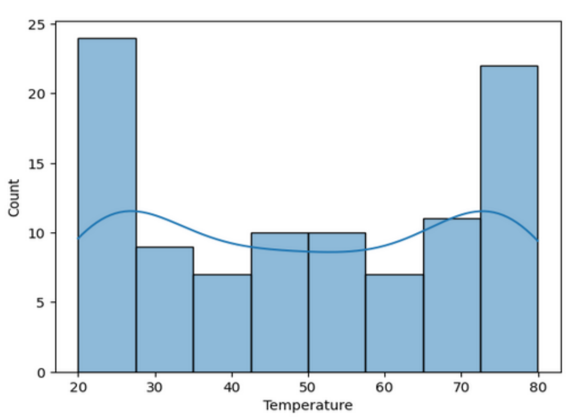

Figure 3: Histogram

Se observa una distribución bimodal, con picos en los extremos cercanos a 20°C y 80°C, indicando una mayor concentración de temperaturas en esos rangos. Las frecuencias en las temperaturas intermedias (30°C - 70°C) son menores. La curva KDE sigue la forma del histograma y confirma la naturaleza bimodal de los datos. Sugiere que las temperaturas tienden a agruparse en torno a dos puntos clave, lo que puede estar relacionado con diferentes condiciones climáticas.

Conclusión: La distribución bimodal implica que un simple modelo lineal podría no ser suficiente para capturar la complejidad de estos datos. Se recomienda considerar esta particularidad al seleccionar un modelo predictivo.

2.3.2. Q-Q Plot

El gráfico Q-Q (Quantile-Quantile) compara los cuantiles de la distribución de los datos observados con los de una distribución teórica normal.

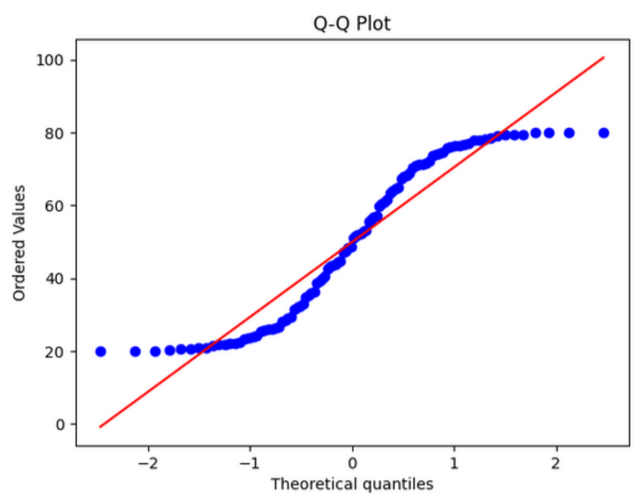

Figure 4: Histogram

1. **Desviación de la línea teórica:** Los puntos observados se desvían de la línea roja de referencia, especialmente en las colas, lo que indica que los datos no siguen una distribución normal.
2. **Forma en 'S':** La forma curva en "S" sugiere una distribución bimodal, confirmando el análisis previo del histograma y la KDE.

Conclusión: El Q-Q Plot demuestra que los datos de temperatura no siguen una distribución normal, lo que implica que los modelos que asumen normalidad podrían no ser adecuados para estos datos.

Metodología

En este estudio se implementarán dos tipos de modelos de regresión para analizar la serie temporal de datos climáticos: uno lineal y uno sinusoidal. Ambos modelos serán evaluados en términos de su capacidad para ajustarse a los datos y realizar predicciones precisas. Para determinar el mejor modelo, utilizaremos el error cuadrático medio (MSE, por sus siglas en inglés) como métrica de evaluación.

- **Modelo de Regresión Lineal:** Este modelo se aplicará tanto sin técnicas de regularización como utilizando las técnicas de regularización L1 (Lasso) y L2 (Ridge). La regularización permite controlar la complejidad del modelo, evitando el sobreajuste, especialmente cuando hay variabilidad en los datos.
- **Modelo Sinusoidal:** Dado que los datos presentan un comportamiento cíclico, se utilizará un modelo de regresión sinusoidal para capturar este patrón. Este modelo será probado tanto sin regularización como con las mismas técnicas de regularización L1 y L2.

La comparación entre ambos modelos se basará en los valores de MSE, con el objetivo de identificar cuál de los dos enfoques proporciona un mejor ajuste a los datos y es más efectivo para predecir futuras mediciones. El modelo con el menor valor de MSE será considerado el más adecuado para este conjunto de datos.

3. Experimentation

3.1. Linear Regression

Reglas de Derivación Utilizadas en la Implementación del Algoritmo

En el algoritmo de regresión lineal con gradiente descendente, se utilizan reglas básicas de derivación para actualizar los parámetros m (pendiente) y c (intersección) de la recta de regresión. A continuación, se muestran las derivadas de la función de costo respecto a m y c.

La función de costo que se minimiza es el **Error Cuadrático Medio (MSE)**, definido como:

$$J(m,c) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + c))^2$$

Donde:

- N es el número de observaciones.
- y_i es el valor real de la variable dependiente.
- x_i es el valor de la variable independiente.
- m es la pendiente (coeficiente de la variable independiente).
- c es la intersección o el término constante.

Derivada de $J(m,c)$ respecto a m

Para encontrar el gradiente de la función de costo respecto a m, derivamos $J(m,c)$ con respecto a m:

$$\frac{\partial J}{\partial m} = \frac{1}{N}\sum_{i=1}^{N} -2x_i(y_i - (mx_i + c))$$

Es decir, el gradiente respecto a m se calcula como:

$$\text{gradient\_m} = -\frac{2}{N}\sum_{i=1}^{N} x_i(y_i - (mx_i + c))$$

Derivada de $J(m,c)$ respecto a c

De manera similar, derivamos $J(m,c)$ con respecto a c:

$$\frac{\partial J}{\partial c} = \frac{1}{N}\sum_{i=1}^{N} -2(y_i - (mx_i + c))$$

Por lo tanto, el gradiente respecto a c se calcula como:

$$\text{gradient\_c} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - (mx_i + c))$$

Funcionamiento de la Regresión Lineal con Gradiente Descendente

La regresión lineal es un método que busca ajustar una línea recta a un conjunto de puntos, de manera que la suma de los errores al cuadrado entre los valores observados Y y los valores predichos por el modelo Y_{pred} sea mínima. Esta línea se define por la ecuación:

$$y = mx + c$$

Donde m es la pendiente de la recta y c es el punto de intersección en el eje y. El objetivo es encontrar los valores óptimos de m y c que minimicen la función de costo (MSE).

Proceso del Algoritmo de Gradiente Descendente

1. **Inicialización**: Comenzamos con valores iniciales de m y c, usualmente 0.

2. **Cálculo del error**: Se calcula el error entre los valores predichos $y_{\text{pred}} = mx + c$ y los valores reales y.

3. **Cálculo de los gradientes**: Se utilizan las derivadas explicadas anteriormente para calcular los gradientes de la función de costo respecto a m y c.

4. **Actualización de los parámetros**: Se actualizan m y c en cada iteración, restando los gradientes multiplicados por la tasa de aprendizaje α:

$$m = m - \alpha \frac{\partial J}{\partial m}$$

$$c = c - \alpha \frac{\partial J}{\partial c}$$

5. **Repetir**: Este proceso se repite durante un número de iteraciones, hasta que los gradientes se acerquen a cero o hasta alcanzar un número máximo de iteraciones. A medida que los gradientes se reducen, los valores de m y c se acercan a los valores óptimos que minimizan el error.

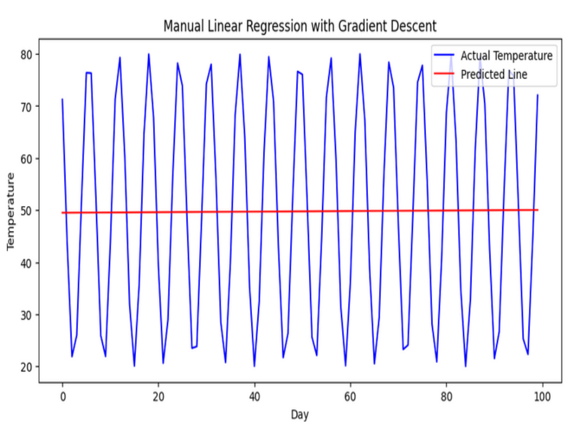

Figure 5: Linear Regression

Conclusión

A partir del análisis del gráfico obtenido y el valor del Error Cuadrático Medio (MSE) de 437.2914, se pueden extraer las siguientes conclusiones:

a) **Desajuste del modelo lineal**: La línea de regresión obtenida es una línea recta horizontal que no captura adecuadamente la variabilidad de los datos, los cuales siguen un patrón oscilante. Esto indica que el modelo lineal no es adecuado para representar la relación entre los días y la temperatura.

b) **Error elevado**: El valor de MSE de 437.2914 es considerablemente alto, lo que refleja que la diferencia entre los valores observados (temperaturas reales) y los valores predichos por el modelo es grande. Esto refuerza la idea de que el modelo lineal no es eficaz para estos datos.

c) **Datos de naturaleza periódica**: El comportamiento observado en los datos muestra una clara oscilación periódica que no puede ser capturada por una regresión lineal. Este tipo de comportamiento cíclico sugiere que una función sinusoidal o un modelo no lineal sería más adecuado.

d) **Necesidad de un modelo no lineal**: Dado que la naturaleza periódica de los datos no puede ser modelada con una regresión lineal, una mejor alternativa sería utilizar un modelo sinusoidal que capture la periodicidad inherente de los datos y proporcione predicciones más precisas.

e) **Conclusión final**: La regresión lineal, al ser un modelo simple y limitado a relaciones lineales, no es apropiada para este conjunto de datos. Se requiere un enfoque más complejo, como un ajuste sinusoidal, para reducir el MSE y proporcionar un ajuste adecuado a los datos observados.

3.2. Sinusoidal Regression

Este algoritmo ajusta un modelo sinusoidal a un conjunto de datos usando el método de gradiente descendente. El modelo sinusoidal sigue la siguiente forma:

$$Y = a \cdot \sin(bX + c) + d$$

Donde a, b, c, y d son los parámetros que el algoritmo ajusta para minimizar el error cuadrático medio (MSE) entre los valores observados y los predichos.

Función del Modelo Sinusoidal

El modelo sinusoidal se define como:

$$\text{sinusoidal}(X, a, b, c, d) = a \cdot \sin(bX + c) + d$$

Donde X es el vector de entradas, a, b, c, y d son los parámetros que ajustaremos.

Función de Costo (MSE)

La función de pérdida utilizada es el Error Cuadrático Medio (MSE), que se define como:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_{\text{actual}} - Y_{\text{predicho}})^2$$

Donde:

- Y_{actual} son los valores observados de los datos.
- Y_{predicho} son los valores predichos por el modelo sinusoidal.

Descenso por Gradiente para Parámetros Sinusoidales

El gradiente descendente se utiliza para minimizar el MSE actualizando los parámetros a, b, c y d. Para ello, se calculan las derivadas parciales del MSE respecto a cada parámetro.

Derivadas Parciales

Las derivadas parciales de la función de costo respecto a cada parámetro son:

$$\frac{\partial MSE}{\partial a} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \cdot \sin(bX_i + c)$$

$$\frac{\partial MSE}{\partial b} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \cdot aX_i \cdot \cos(bX_i + c)$$

$$\frac{\partial MSE}{\partial c} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \cdot a \cdot \cos(bX_i + c)$$

$$\frac{\partial MSE}{\partial d} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)$$

Donde:

- n es el número de puntos de datos.
- X_i es el valor de la variable independiente.
- Y_i es el valor observado.
- $\hat{Y}_i$ es el valor predicho por el modelo.

Actualización de Parámetros

Los parámetros se actualizan en cada iteración usando la siguiente fórmula:

$$a = a - \alpha \cdot \frac{\partial MSE}{\partial a}$$

$$b = b - \alpha \cdot \frac{\partial MSE}{\partial b}$$

$$c = c - \alpha \cdot \frac{\partial MSE}{\partial c}$$

$$d = d - \alpha \cdot \frac{\partial MSE}{\partial d}$$

Donde α es la tasa de aprendizaje.

Implementación del Gradiente Descendente

El algoritmo de gradiente descendente ajusta los parámetros a, b, c, y d minimizando el MSE. En cada iteración, los parámetros se actualizan en función de los gradientes.

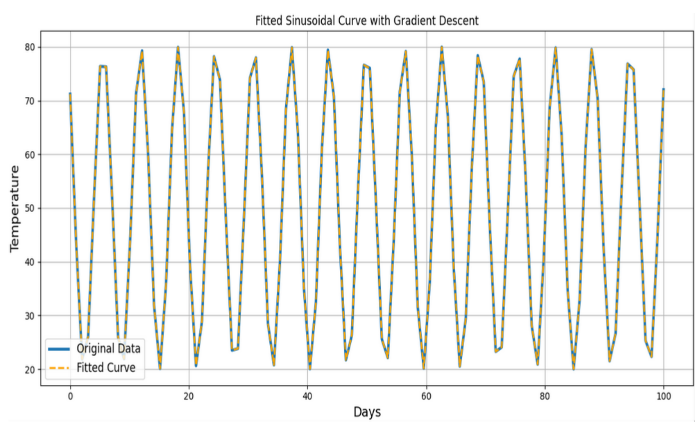

Figure 6: Sinusoidal Regression

Conclusión

A partir del ajuste sinusoidal realizado sobre los datos de temperatura y el gráfico obtenido, se pueden extraer las siguientes conclusiones:

a) **Buen ajuste del modelo sinusoidal**: El modelo sinusoidal ajustado (línea discontinua naranja) sigue muy de cerca los valores reales de los datos (línea azul). Esto indica que el modelo sinusoidal ha capturado con éxito el patrón periódico de los datos.

b) **Captura de la periodicidad**: El ajuste sinusoidal refleja con precisión la naturaleza cíclica de los datos. Los máximos y mínimos de la curva ajustada coinciden bien con los de los datos originales, lo que sugiere que el modelo ha logrado modelar correctamente la frecuencia de los ciclos.

c) **Convergencia con gradiente descendente**: El algoritmo de gradiente descendente ha permitido encontrar los parámetros óptimos a, b, c, y d que mejor describen los datos. A pesar de haber utilizado un número elevado de iteraciones (epochs) y una tasa de aprendizaje baja, el algoritmo ha convergido hacia una solución precisa, minimizando el error cuadrático medio (MSE).

d) **Bajo error cuadrático medio (MSE)**: El valor del MSE se ha reducido significativamente durante el proceso de gradiente descendente, lo que confirma que el modelo ajustado es adecuado para los datos. Un bajo MSE indica que la diferencia entre los valores reales y los predichos es mínima.

e) **Recomendación de uso del modelo sinusoidal**: Dado el comportamiento cíclico claro en los datos de temperatura, la elección de un modelo sinusoidal es apropiada y superior a un ajuste lineal, que no habría podido capturar la oscilación periódica observada.

Sinusoidal Regression Regularization

L1 Regularization

La **regularización L1**, también conocida como **Lasso** (*Least Absolute Shrinkage and Selection Operator*), es una técnica que agrega un término de penalización basado en la *suma de los valores absolutos* de los coeficientes del modelo. El objetivo de esta regularización es evitar el sobreajuste del modelo y hacer que algunos coeficientes se reduzcan a cero, lo que conduce a una mayor simplicidad del modelo.

Fórmula de la Regularización L1:

$$\text{Costo total} = \text{MSE} + \alpha_{L1} \sum_i |w_i|$$

Donde:

- MSE es el *Error Cuadrático Medio*, que mide la diferencia entre los valores predichos y los valores reales.
- α_{L1} es el hiperparámetro de regularización, que controla la magnitud de la penalización.
- w_i son los coeficientes del modelo.

El término $\alpha_{L1} \sum_i |w_i|$ se conoce como el **término de regularización L1**, y su efecto principal es que reduce algunos de los coeficientes del modelo a cero cuando la penalización es alta, lo que induce un modelo más esparcido o parsimonioso.

Aplicación en el Código: En el código, la regularización L1 se incorpora al modificar la función de pérdida (*loss*) con un término de penalización. El término de regularización α_{L1} se añade a la pérdida *MSE*, y afecta el valor de los gradientes de los parámetros.

Los coeficientes a, b, c y d son ajustados en cada iteración usando gradiente descendente, con un término adicional que reduce su magnitud en función de α_{L1} y el signo del parámetro.

$$\text{MSE}_{L1} = \text{MSE} + \alpha_{L1} \left(|a| + |b| + |c| + |d|\right)$$

Las derivadas parciales de la función de costo con regularización L1 respecto a cada parámetro son:

$$\frac{\partial \text{MSE}_{L1}}{\partial a} = \frac{\partial \text{MSE}}{\partial a} + \alpha_{L1} \cdot \text{sign}(a)$$

$$\frac{\partial \text{MSE}_{L1}}{\partial b} = \frac{\partial \text{MSE}}{\partial b} + \alpha_{L1} \cdot \text{sign}(b)$$

$$\frac{\partial \text{MSE}_{L1}}{\partial c} = \frac{\partial \text{MSE}}{\partial c} + \alpha_{L1} \cdot \text{sign}(c)$$

$$\frac{\partial \text{MSE}_{L1}}{\partial d} = \frac{\partial \text{MSE}}{\partial d} + \alpha_{L1} \cdot \text{sign}(d)$$

Donde:

- α_{L1} es el hiperparámetro de regularización L1.
- $\text{sign}(x)$ es la función signo, que devuelve 1 si $x > 0$, -1 si $x < 0$, y 0 si $x = 0$.

L2 Regularization

La **regularización L2**, también conocida como **Ridge**, agrega un término de penalización basado en la *suma de los cuadrados* de los coeficientes del modelo. A diferencia de L1, L2 no fuerza los coeficientes a ser cero, sino que los reduce, estabilizando el modelo y previniendo el sobreajuste.

Fórmula de la Regularización L2:

$$\text{Costo total} = \text{MSE} + \alpha_{L2} \sum_i w_i^2$$

Donde:

- MSE es el *Error Cuadrático Medio.*
- α_{L2} es el hiperparámetro de regularización L2, que controla el peso de la penalización.
- w_i son los coeficientes del modelo.

El término de regularización L2 $\alpha_{L2} \sum_i w_i^2$ penaliza los grandes valores de los coeficientes y tiende a reducir su magnitud, lo que hace que el modelo sea más suave. Sin embargo, a diferencia de L1, no lleva los coeficientes a cero, lo que mantiene todas las características activas.

Aplicación en el Código: Al igual que en L1, la regularización L2 se añade a la función de pérdida y afecta los gradientes de los coeficientes. Los coeficientes a, b, c y d son ajustados con un término adicional $2\alpha_{L2} w_i$, que reduce la magnitud de los coeficientes de manera proporcional a su valor actual. Ambas regularizaciones se aplican durante el entrenamiento del modelo sinusoidal mediante el método de gradiente descendente.

$$\text{MSE}_{L2} = \text{MSE} + \alpha_{L2}\left(a^2 + b^2 + c^2 + d^2\right)$$

Las derivadas parciales de la función de costo con regularización L2 respecto a cada parámetro son:

$$\frac{\partial \text{MSE}_{L2}}{\partial a} = \frac{\partial \text{MSE}}{\partial a} + 2\alpha_{L2} a$$

$$\frac{\partial \text{MSE}_{L2}}{\partial b} = \frac{\partial \text{MSE}}{\partial b} + 2\alpha_{L2} b$$

$$\frac{\partial \text{MSE}_{L2}}{\partial c} = \frac{\partial \text{MSE}}{\partial c} + 2\alpha_{L2} c$$

$$\frac{\partial \text{MSE}_{L2}}{\partial d} = \frac{\partial \text{MSE}}{\partial d} + 2\alpha_{L2} d$$

Donde:

- α_{L2} es el hiperparámetro de regularización L2.

Gráfica de comparación

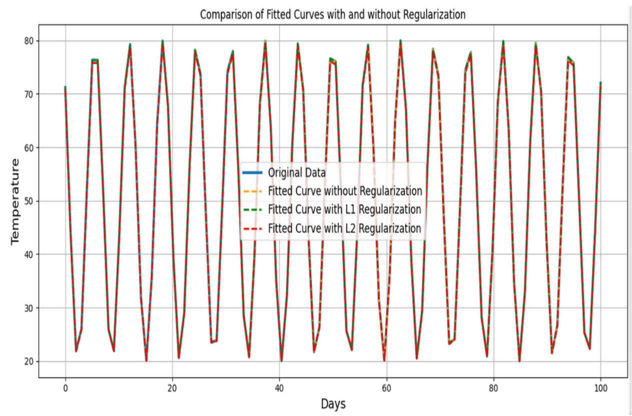

Figure 7: Sinusoidal Regression Regularization

4. Discussion

Relación entre el código de L1/L2 y Machine Learning

El código presentado aplica regularización L1 y L2 en un modelo de regresión sinusoidal con descenso de gradiente. A continuación, se relacionan algunos conceptos teóricos relevantes con lo que ocurre en el código.

Sin Regularización

El código ajusta un modelo sinusoidal a los datos de temperatura sin aplicar regularización, lo que implica que el modelo tiene la libertad total para ajustarse a los datos de entrenamiento. Este ajuste puede llevar a un **overfitting**, ya que el modelo tiene la capacidad de capturar toda la varianza en los datos de entrenamiento, pero puede no generalizar bien a nuevos datos. Esto se refleja en un error cuadrático medio (**MSE**) bajo en el entrenamiento, pero posiblemente alto en nuevos datos.

Con Regularización L1 y L2

El código luego introduce regularización L1 (Lasso) y L2 (Ridge) al modelo sinusoidal. La regularización penaliza los parámetros del modelo, forzando al modelo a ser más simple y reduciendo su capacidad de ajuste exacto a los datos. Esto reduce el riesgo de **overfitting**, pero incrementa el MSE en los datos de entrenamiento, lo que indica una menor precisión en los datos ajustados.

Bias-Variance Tradeoff

La regularización afecta directamente el **bias-variance tradeoff**. Sin regularización, el modelo tiene baja varianza y alto sesgo, lo que puede llevar al sobreajuste. Al aplicar regularización L1 o L2, el sesgo aumenta (el modelo es más rígido), pero la varianza se reduce, lo que mejora la capacidad del modelo para generalizar a datos nuevos. El reto es encontrar el punto óptimo de regularización donde el MSE en los datos no vistos sea minimizado.

Overfitting y Underfitting

Este equilibrio es clave para evitar tanto **overfitting** (donde el modelo es demasiado flexible y ajusta el ruido en los datos de entrenamiento) como **underfitting** (donde el modelo es demasiado simple y no captura la relación subyacente en los datos). La regularización ayuda a evitar el sobreajuste, pero si la penalización es demasiado fuerte, puede llevar al subajuste.

5. Conclusions

Limitaciones del código:

- Tasa de aprendizaje baja: En el código, la tasa de aprendizaje (`learning_rate`) es extremadamente baja (10^{-7}). Esto hará que el modelo tarde mucho en converger hacia los valores óptimos de los parámetros, lo que alarga innecesariamente el proceso de entrenamiento.
- Número de épocas elevado: El código tiene $5,000,000$ épocas, lo cual es un número excesivamente alto, especialmente considerando la baja tasa de aprendizaje. Este valor podría ser ineficiente y consumir mucho tiempo y recursos sin aportar beneficios significativos si el modelo ya ha convergido mucho antes.

Recomendaciones:

1. Ajustar la tasa de aprendizaje: Se podría experimentar con una tasa de aprendizaje más alta, por ejemplo, entre 10^{-3} y 10^{-5}, para lograr una convergencia más rápida sin sacrificar precisión. Esto permitiría al modelo ajustarse a los datos de manera más eficiente y reducir significativamente el tiempo de entrenamiento.
2. Reducir el número de épocas: Se debe reducir el número de épocas a un valor más razonable, como $100,000$ o menos, y monitorear el error de validación para detectar el momento en que el modelo deja de mejorar. Esto evitará el entrenamiento innecesario y ahorrará recursos computacionales.
3. Implementar Elastic Net: Combinar las regularizaciones L1 y L2 con Elastic Net puede ofrecer mejores resultados, especialmente si el objetivo es tener un modelo más robusto y equilibrado en términos de sobreajuste y subajuste. Elastic Net combina las ventajas de Lasso y Ridge, ayudando a manejar tanto la selección de características como la reducción de varianza.

Resultados:

1. Se obtuvo un MSE de 0.00021639969754827577 para el caso sin regularización. El Error Cuadrático Medio (MSE) en el modelo sin regularización es muy bajo, lo que indica que el modelo se ajusta bien a los datos de entrenamiento. Sin embargo, un MSE tan bajo podría sugerir que el modelo está sobreajustado (overfitting).
2. Al aplicar regularización L1, el MSE aumentó a 3.867006023816327 usando un alpha de 0.04641588833612782. Este aumento es debido a que al penalizar los coeficientes, algunos de ellos pueden volverse exactamente cero, simplificando el modelo y haciéndolo más robusto ante el sobreajuste. Sin embargo, el alto MSE en este caso sugiere que la magnitud de la regularización es demasiado alta, lo que podría estar llevando a un subajuste (underfitting).
3. Al aplicar L2, el MSE también aumenta, pero no tanto como en el caso de la regularización L1. Aumentó a 1.7884308688946173 con un alpha de 0.01. Este aumento se debe a que está siendo penalizado, mantiene todos los coeficientes del modelo, pero los reduce en magnitud para evitar un sobreajuste. Este resultado sugiere que el modelo está mejor equilibrado entre ajuste y generalización en comparación con la regularización L1.
4. El modelo sin regularización tiene el menor MSE, lo que indica un buen ajuste a los datos de entrenamiento, pero corre el riesgo de sobreajustarse. La regularización L1 penaliza fuertemente

el modelo, llevándolo posiblemente a un subajuste, pero a cambio genera un modelo más simple y posiblemente más robusto para datos nuevos. La regularización L2 ofrece un término medio, proporcionando un aumento en el MSE, pero con menor riesgo de sobreajuste en comparación con el modelo sin regularización y sin simplificar tanto como el modelo L1.

6. Citations and Bibliography

[1] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

7. Source Code

Repo