Métodos Numéricos

30 de junio de 2024

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Trabajo práctico 3

Primer cuatrimestre 2024

Grupo 19

Integrante	LU	Correo electrónico
Rios, Brian Ivan	917/19	ivan.rios2010@gmail.com
Saccomano, Ignacio Manuel	222/22	nachosacco1@gmail.com
Maldonado, Juan Bautista	164/22	jbmaldonado2003@gmail.com

Resumen

Estudio de la técnica de cuadrados mínimos para la correcta aproximación de un conjunto de datos a partir de los polinomios de Legendre en lugar de una base de polinomios fija. Posteriormente, se empleará la optimización de hiperparámetros a través de regularización y validación cruzada para garantizar una solución precisa y efectiva para datos desconocidos construida a partir de datos de ajuste.

Palabras clave: Cuadrados mínimos, regularización, SVD, interpolación de Legendre.

Índice

1. Introducción

En diversas áreas del conocimiento, es un problema recurrente el ajuste de modelos a conjuntos de datos desconocidos a partir de una muestra conocida. Esto es útil en la mayoría de aplicaciones modernas, donde muchos fenómenos siguen una distribución desconocida o, en su defecto, se debe hallar una aproximación porque la función asociada a la distribución subyacente es demasiado compleja para usarse en la práctica.

Para llegar a la aproximación de esta distribución se elige una muestra significativa de datos y se pueden usar para ajustar los parámetros de una función $f(x)$ conocida que permita aproximar la distribución inicial. Es posible usar diversos criterios para optimizar la aproximación, y en este informe la vamos a medir en función de la suma de cuadrados de las diferencias entre los puntos generados por la función elegida y los correspondientes valores en los datos.

Para construir $f(x)$ se puede utilizar una base de polinomios linealmente independiente. A esto se lo conoce como regresión polinomial [1]. Dependiendo de los polinomios que se elija para construir la función, sin embargo, se puede incurrir en error debido al numéro de condición de la matriz resultante. Para minimizar este problema se pueden emplear los **polinomios de Legendre** que además son una base ortonormal [2], lo que otorgará ciertas propiedades útiles como se verá posteriormente.

Agregado a lo previamente mencionado, en diversos contextos se agregan otros requerimientos para la solución que se busca. Por ejemplo, probará ser útil pedir que los coeficientes a encontrar no sean arbitrariamente grandes lo que otorgará soluciones menos propensas a errores. Al agregado de requerimientos se le conoce como **regularización**, y abordaremos específicamente la **L2** o regresión *ridge* [3].

En este trabajo, entonces, se estudiará el problema de la regresión con polinomial usando el criterio de aproximación de cuadrados mínimos lineales. En las secciones siguientes se presentará un desarrollo detallado del marco teórico del problema mencionado, en conjunto con el uso de los polinomios de Legendre y la descomposición en valores singulares para la correcta aproximación a un conjunto de datos.

Sumado a ello, se abordará la optimización de hiperparámetros mediante validación cruzada, con el objetivo de mejorar la precisión y generalización del modelo de una forma eficiente.

2. Desarrollo

2.1. Cuadrados mínimos

El objetivo principal de cuadrados mínimos es encontrar una función que ajuste un conjunto de datos de manera óptima. Esto implica minimizar la discrepancia entre los valores observados y los valores predichos por el modelo. La técnica se basa en minimizar la suma de los cuadrados de las diferencias entre los valores observados y_i y los predichos $\hat{y_i}$ por la función $f(x_i)$. Si se tiene un conjunto de datos (x_i, y_i) para i = 1, 2,..., n y una función $f(x)$ perteneciente a una familia de funciones F tal que mejor aproxime a los datos, formalmente el problema se define como:

$$\min_{f \in F} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

Dado que la función f no siempre es lineal, para el problema de cuadrados mínimos lineales se debe descomponer a f en una familia de funciones F de modo tal que resulte en una combinación lineal de ellas. Es decir $f(x) = \sum_{j=1}^{n} \beta_j \phi_j(x)$ donde $\phi_j(x) \in F$. Reescribiendo el problema inicial:

$$\min_{\beta_1,\ldots,\beta_n} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \beta_j \phi_j(x_i) - y_i \right)^2$$

Dado que la función f fue expresada como combinación lineal de $\phi_j(x)$ podemos definir $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, y $x \in \mathbb{R}^n$ como [1]:

$$A = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_m) & \phi_2(x_m) & \cdots & \phi_n(x_m) \end{bmatrix} \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad x = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

Se puede expresar el problema de cuadrados mínimos en forma matricial como:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$

El problema inicial se redujo a minimizar la resta entre 2 vectores. Si Ax es igual a y, entonces minimizamos la expresión puesto que la norma 2 al cuadrado de un vector como mínimo puede ser 0. Si esto no se cumple, para minimizar la expresión tiene que ocurrir que Ax es la proyección ortogonal sobre $Im(A)$ de b [1]. El vector resultado de la minimización se encuentra en $Im(A)^{\perp}$ la cual es igual al $Nu(A^T)$, por lo que:

$$\begin{aligned} A^T(Ax - b) &= 0 \\ A^T Ax - A^T b &= 0 \\ (A^T A)x &= A^T b \end{aligned}$$

Estas son las ecuaciones normales y con $A^T A$ inversible la solución al sistema puede despejarse como:

$$x = (A^T A)^{-1} A^T b \qquad (2.1)$$

Puesto que el cómputo de la inversa de una matriz es costoso, será de interés calcular la descomposición en valores singulares *(SVD)* de la matriz A. Se define la descomposición SVD de A como sea $A \in \mathbb{R}^{m \times n}$, donde r = rango(A). Existen $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ matrices ortogonales, y $\Sigma \in \mathbb{R}^{m \times n}$ tal que: $A = U\Sigma V^T$ y Σ es una "diagonal" con valores positivos ordenados. Si se considera el caso A tiene rango completo, la descomposición SVD puede reducirse para que Σ, U y V compartan dimensiones las filas nulas de Σ y las columnas de V o U dependiendo si $m > n$ o viceversa. Esto es lo que se conoce como *economy size SVD*[5]. De este modo se obtiene:

$$x = V\Sigma^{-1} U^T b$$

Dado que ahora A está definida por una matriz diagonal y dos matrices ortogonales, el cómputo de la inversa es significativamente más eficiente. Solo basta con trasponer a U y V e invertir los coeficientes de la diagonal de Σ.

Entonces, encontrar el β que resuelve el problema de cuadrados mínimos para una matriz X y un vector y se reduce a las siguientes instrucciones:

Algorithm 1 Solución a cuadrados mínimos lineales

```
1: procedure CML(X, y)
2:     m ← #cols(X)
3:     n ← #filas(X)
4:     X ← UΣV^t                                  ▷ La descomposición SVD de X
5:     if m > n then
6:         U ← U[:, : n]                          ▷ U ahora solo contiene sus primeras n columnas
7:         Σ ← Σ[:n, :n]                          ▷ Σ ahora contiene sus primeras n filas y n columnas
8:     else if m < n then
9:         V ← V[:, : m]                          ▷ V ahora solo contiene sus primeras m columnas
10:        Σ ← Σ[:m, :m]                          ▷ Σ ahora contiene sus primeras m filas y m columnas
11:    end if
12:    β ← VΣ^{-1}U^t y
13:    return β
14: end procedure
```

2.2. Regresión polinomial con polinomios de Legendre

Resulta fundamental mantener el número de condición de la matriz resultante bajo. Caso contrario, pequeñas perturbaciones pueden llevar a soluciones inestables y amplificación del error a magnitudes intolerables como puede ocurrir si se elige la base $\{1, x, x^2, ..., x^n\}$.

Los polinomios de Legendre [2] constituyen una base ortonormal por lo que la matriz asociada tiene un número de condición bajo, permitiendo así mantener un grado de precisión tolerable ante los problemas numéricos asociados a la aritmética finita. Por lo tanto, al usarlos como base se reduce la amplificación del error y se obtienen soluciones más robustas en el problema de cuadrados mínimos y por ser además inversible se puede realizar el despeje de β mostrado previamente en la ecuación (2.1) y el posterior despeje mostrado en la ecuación (2.3). Sumado a la facilidad que proporciona la descomposición SVD para el cómputo de la inversa, esto permite calcular fácilmente los coeficientes a despejar.

Ahora tomamos un conjunto de datos con una distribución desconocida que buscaremos aproximar con los polinomios de Legendre. Este conjunto se particionará en un subconjunto de ajuste y otro de validación para aplicar ***cross-validation***. Lo que haremos será aplicar regresión tomando solamente los datos de ajuste, construyendo con ellos la matriz asociada a los polinomios de Legendre que notaremos X_{ajuste}. Finalmente, veremos qué relación hay entre el grado de los polinomios que se usen para aproximar los datos y el error cuadrático medio (**ECM**) entre los datos predichos $\hat{y}$ para los de ajuste y validación respectivamente.

Formalmente, sea $\beta = \hat{X}_{ajuste} y_{ajuste}$ donde $\hat{X}_{ajuste}$ es la matriz de los polinomios de Legendre de grado 1 hasta i de los datos de ajuste. En cada iteración calcularemos el error cuadrático medio de $\hat{X}_{ajuste}\beta - y_{ajuste}$ y de $\hat{X}_{val}\beta - y_{val}$ donde $\hat{X}_{val}$ es la matriz de los polinomios de Legendre de grado 1 hasta i construida a partir de los datos de validación. Con esto último queremos ver qué tanto se relaciona la función de aproximación de los datos de ajuste a los de validación o, lo que es lo mismo, si la distribución subyacente del fenómeno está bien explicada por los polinomios.

Para la experimentación de este último punto se consideró el polinomio de Legendre de grados 1 hasta el grado $2n$ donde n es el tamaño de la muestra de datos de ajuste.

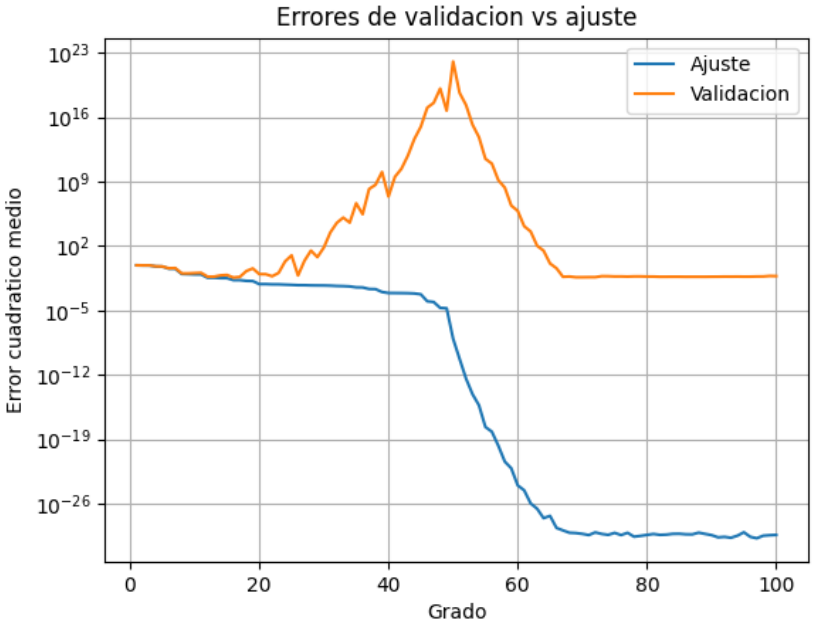

Figura 1: Comparación del *ECM* de los datos de ajuste y validación a medida que aumenta el grado del polinomio. Como se puede observar aproximadamente en el grado 50 se observa el fenómeno de doble descenso [4].

Como se puede observar en la figura 1 el ECM disminuye paulatinamente a medida que el grado de los polinomios aumenta. Esto es debido a que, al aumentar el grado, la función de aproximación se acerca cada vez más a la función que determina la distribución de la muestra. Sin embargo, alrededor del grado 50 se puede observar que a pesar de que el ECM de los datos de ajuste disminuyen, el error se dispara para datos desconocidos. Este fenómeno puede explicarse porque a pesar de que los datos estén gobernados por una distribución desconocida algunos puntos caen fuera de ella y al aumentar el grado se les da mucha relevancia provocando que el modelo prediga erróneamente valores nuevos por asignarle demasiado peso a puntos que no debería. A esto se lo conoce como **sobreajuste**, y ocurre cuando la cantidad de parámetros del modelo (en este caso el grado de los polinomios) coincide con el tamaño de la muestra. A medida que la cantidad de parámetros aumenta el modelo tiene más información para discriminar los *outliers* y mejorando así la efectividad.

A pesar de que el error de los datos de ajuste se redujo a niveles despreciables con grado 100, el error de los datos de validación es mucho mayor y no mejora con el grado luego del descenso. Para mejorar esto se puede pedir más condiciones a la solución final.

2.3. Cuadrados mínimos con regularización

El sobreajuste es un error común donde un modelo se adapta excesivamente a los datos de entrenamiento, lo que resulta en una mala generalización de las soluciones calculadas. Un modelo complejo puede ajustar muy bien los datos de entrenamiento pero puede tener un rendimiento pobre en datos nuevos debido al sobreajuste. La regularización introduce un término de penalización en la función objetivo que controla la magnitud de los coeficientes del modelo, forzándolos a ser más pequeños.

Como fue mencionado se utilizará la regularización L2, también conocida como regresión *ridge*. Esta añade un término de penalización para los coeficientes grandes al sumar $||\beta||_2^2$ multiplicado por un escalar λ que debe ajustarse dependiendo del grado de penalización deseado. Formalmente el problema de cuadrados mínimos con regularización ahora se define como:

$$\min_{\beta_1,\ldots,\beta_n} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \beta_j \phi_j(x_i) - y_i \right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

Con este agregado las ecuaciones normales previamente descritas ahora se definen de la siguiente manera:

$$(A^T A + \lambda I)\beta = A^T y$$

Con $(A^T A + \lambda I)$ inversible se consigue:

$$\beta = (A^T A + \lambda I)^{-1} A^T y \qquad (2.3)$$

Anteriormente se mencionó que utilizar la descomposición en valores singulares de A resulta muy beneficioso a nivel computacional. Para la exploración de error e hiperparámetros se define el problema de cuadrados mínimos con regularización y SVD como:

$$\beta = (A^T A + \lambda I)^{-1} A^T y$$

$$\beta = ((U\Sigma V^T)^T (U\Sigma V^T) + \lambda I)^{-1} (U\Sigma V^T)^T y$$

$$\beta = (V\Sigma^2 V^T + \lambda I)^{-1} V\Sigma U^T y$$

$$\beta = (V\Sigma^2 V^T + \lambda V V^T)^{-1} V\Sigma U^T y$$

$$\beta = (V(\Sigma^2 + \lambda I)V^T)^{-1} V\Sigma U^T y$$

$$\beta = V(\Sigma^2 + \lambda I)^{-1} V^T V\Sigma U^T y$$

$$\beta = V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y \qquad \triangleright \text{ Producto de diagonales es simétrico}$$

$$\boxed{\beta = V\Sigma(\Sigma^2 + \lambda I)^{-1} U^T y}$$

Donde el cálculo de la inversa no es costoso pues es otra diagonal compuesta por sus inversos multiplicativos. De este modo, ahora β se calcula de la siguiente manera:

Algorithm 2 Solución a cuadrados mínimos lineales con regularización

```
procedure CML_Reg(X, y, λ)
    m ← #cols(X)
    n ← #filas(X)
    X ← UΣV^t                                   ▷ La descomposición SVD de X
    if m > n then
        U ← U[:, : n]                           ▷ U ahora solo contiene sus primeras n columnas
        Σ ← Σ[:n, :n]                           ▷ Σ ahora contiene sus primeras n filas y n columnas
    else if m < n then
        V ← V[:, : m]                           ▷ V ahora solo contiene sus primeras m columnas
        Σ ← Σ[:m, :m]                           ▷ Σ ahora contiene sus primeras m filas y m columnas
    end if
    β ← VΣ(Σ^2 + λI)^{-1} U^T y
    return β
end procedure
```

Con esta nueva restricción para β ahora introducimos otra experimentación donde además de explorar diversos grados para X_{ajuste} probaremos distintos coeficientes de penalización λ para ver cómo se comporta el error.

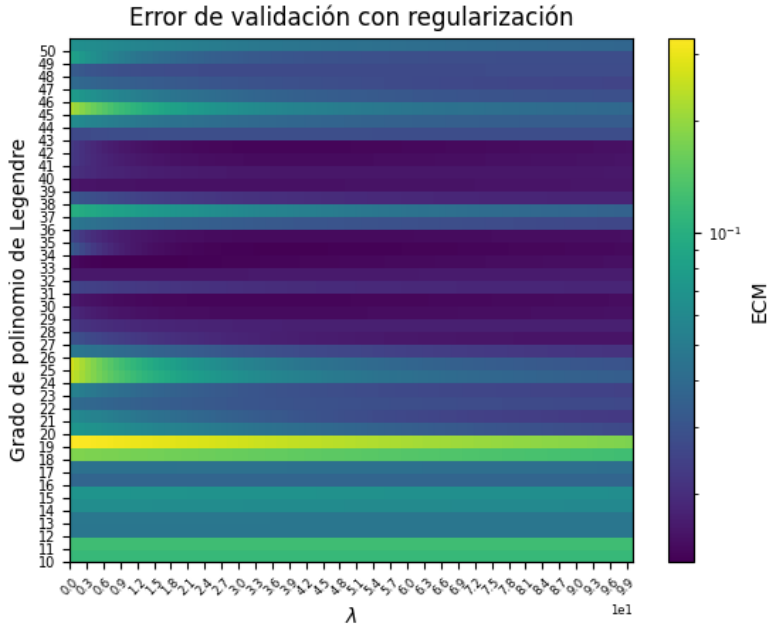

Figura 2: Mapa de calor del error respecto a los datos de validación para cada combinación de grado de polinomio y coeficiente de penalización, tomando los polinomios entre grado 10 y 50 y 100 muestras de λ entre $0,01$ y $0,1$. Las abscisas están expresadas en notación científica y la temperatura está representada de forma logarítmica para facilitar la visualización del 'valle' donde se minimiza el error para el λ y grado óptimos.

Luego de haber analizado el error para diversos coeficientes λ se realizó un acercamiento a la región que ocupa el mínimo, comprendida por la **figura 2.3**. Como se puede observar el error decrementa a medida que aumenta el grado, alcanza el mínimo en el grado 33 y $\lambda \approx 0,02$ y luego del mínimo el error vuelve a incrementar. Esto es congruente con el comportamiento esperado del mínimo, pues la región aledaña al dominio termina en valores crecientes a medida que se alejan del punto.

Este crecimiento cercano al mínimo además expresa la relevancia del coeficiente de penalización, pues al desplazarse levemente del óptimo permite que los coeficientes β alcancen valores que alejan a las predicciones del valor objetivo.

3. Conclusiones

Se exploró el uso de los polinomios de Legendre para el método de regresión en el contexto del modelado de problemas de adaptación a datos desconocidos a partir de otros conocidos. Para calcular la aproximación de los datos predichos a los reales se utilizó el criterio de cuadrados mínimos lineales y se exploró el error cuadrático medio para un conjunto de hiperparámetros.

Los polinomos de Legendre probaron ser una gran herramienta por las propiedades de la matriz resultante, en particular el hecho de que sea una base ortonormal que garantiza un número de condición mínimo y la existencia de una inversa que permita despejar fácilmente los coeficientes a calcular. Adicionalmente, la descomposición SVD de la matriz asociada probó agilizar el cómputo de la inversa que es una condición necesaria para modelos que involucran una gran cantidad datos y parámetros como suele ocurrir en la mayoría de aplicaciones.

A pesar de la optimalidad de la solución, quedó evidenciado el fenómeno de doble descenso en el grado coincidente con la cantidad de la muestra demostrando que se debe tener resguardo a la hora de elegir la cantidad de parámetros para la regresión.

En vista de ello se propuso la técnica de regularización con la implementación de un coeficiente de penalización que mostró reducir el error de los datos de validación al evitar el fenómeno de sobreajuste. Este agregado influye más a medida que crece la cantidad de parámetros, siendo imprescindible para modelos donde la presencia de outliers condicione excesivamente la efectividad de los mismos.

4. Bibliografía

Referencias

[1] Clases teóricas

[2] Información del enunciado del TP

[3] Diapositivas del laboratorio

[4] Fenómeno de doble descenso

[5] Economy size SVD

5. Apéndice

Para el desarrollo de los experimentos se utilizó el lenguaje **Python** junto con las bibliotecas de **Numpy** para los cálculos matriciales y **Pyplot** para los gráficos.