Optimización de Aprendizaje automático de una IA con Álgebra Lineal: Abordando la Maldición de la Dimensionalidad con Regresión lineal (algoritmo supervisado)

Tiffany A. Jordán-Uquillas, Steffanie D. C. Alvarado, Kevin E. Valverde-Mullo y Erick A. Guarnizo-Carrera

Universidad Tecnológica Ecotec

Algebra lineal

Mgtr. Giraldo De La Caridad Leon Rodriguez

13 de diciembre del 2023

Introducción

Problemática: La alta dimensionalidad de los datos es un obstáculo común en el aprendizaje automático, lo que resulta en modelos ineficientes, sobreajuste y tiempos de entrenamiento prolongados. La cuestión principal es cómo reducir la dimensionalidad y conservar características relevantes para optimizar la IA.

Modelo: Combinación lineal, matriz transpuesta, reducción de matriz y matriz inversa.

Posible solución: La solución propuesta consiste en aplicar técnicas de reducción de dimensionalidad supervisada de la regresión lineal.

Definiciones

Aprendizaje automático: Es un subcampo de la inteligencia artificial (IA) que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones a partir de datos sin ser explícitamente programadas.

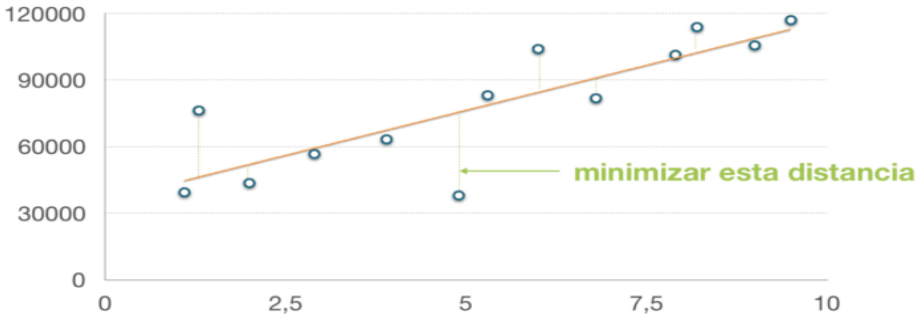
Alta dimensionalidad: Conjuntos de datos que tienen un gran número de características o variables. Como consecuencia el procesamiento y el almacenamiento de datos de ese tipo son mucho más costosos en términos computacionales, ya que los cálculos se vuelven más lentos, y la memoria requerida aumenta considerablemente.

Regresión lineal (técnica de aprendizaje supervisado): Es un método estadístico que se utiliza para modelar la relación entre una variable dependiente (también llamada variable de respuesta) y una o más variables independientes (también llamadas predictores o variables explicativas).

- La regresión lineal simple, tiene una variable independiente y una variable dependiente. La ecuación de la línea se expresa como:

$$Y = mx + b$$

    - y es la variable dependiente.
    - x es la variable independiente.
    - m es la pendiente de la línea.
    - b es la ordenada al origen (el valor de Y cuando X es igual a cero)
- La Regresión Lineal Simple busca minimizar la distancia vertical entre todos los datos y nuestra línea, por lo tanto, para determinar la mejor línea, debemos minimizar la distancia entre todos los puntos y la distancia de nuestra línea.



- La regresión múltiple es donde hay más de una variable independiente, en este caso es la que usaremos ya que se abordará la maldición de la dimensionalidad.

$$Y= b0 +b1x1 + b2x2+…+bnXn$$

  - y es la variable dependiente
  - b0 es el término de intersección (constante)
  - b1,b2,…,bn son los coeficientes asociados con las variables independientes
  - x1,x2,...,xn.

<u>Tipos principales de aprendizaje automático</u>

- **Aprendizaje Supervisado:** El modelo aprende en relación a los datos que ingresan, de acuerdo a esos datos hace predicciones de datos que aún no sabe.
- **Aprendizaje No Supervisado:** El modelo debe encontrar patrones o estructuras en los datos por sí mismo.

<u>Caso de estudio</u>

Paso 1: Definir el Problema

PropTech Solutions, una empresa líder en el análisis de datos inmobiliarios. Cuenta con un conjunto de datos que incluye información sobre los precios de los departamentos de la ciudad de Guayaquil en el sector norte. La empresa desea desarrollar un modelo predictivo que les permita ofrecer evaluaciones precisas sobre los precios de los departamentos a sus clientes.

Supongamos que la información de precio (Y) con la que cuenta la empresa está en función de cuatro características: dormitorios (A), baños (B), lavandería (Si tiene o no) (C), estacionamiento (si tiene o no) (D). Queremos determinar cuáles de estas características son las más importantes para predecir el precio de los departamentos.

Paso 2: Preparar los Datos

Tenemos los datos de precios de los departamentos y las cuatro características (A, B, C, D). Los datos se ven así:

Dormitorios (A)	Baños (B)	Lavandería (Si tiene o no) (C)	Estacionamiento (Si tiene o no) (D)	Precio por departamento
2	1	1	0	150
3	2	0	1	200
4	3	1	1	250
4	1	1	0	170
3	2	1	1	220
2	1	0	1	180
4	2	1	0	210

Paso 3: Aplicar Regresión Lineal

Utilizaremos una regresión lineal múltiple para encontrar los coeficientes que relacionan las características con el precio de las viviendas. La relación se expresa como:

Precio = b0 + b1 * A + b2 * B + b3 * C + b4 * D + b5 * E

Donde b0 es el término de sesgo (intercept) y b1, b2, b3, b4 y b5 son los coeficientes (valores) de las características A, B, C, D y E, respectivamente.

- **Sesgo:** Se refiere a la diferencia entre las predicciones del modelo y los valores reales en los datos de entrenamiento. En otras palabras, es una medida de cuán lejos está la predicción del modelo del resultado deseado.

  Un sesgo alto implica que el modelo es demasiado simple y no puede capturar la complejidad subyacente de los datos, lo que conduce a predicciones inexactas. Por otro lado, un sesgo bajo indica que el modelo es lo suficientemente complejo como para adaptarse a los datos de entrenamiento, pero puede correr el riesgo de sobreajustarse y no generalizar bien a nuevos datos.

El valor de 1 se usa en el sesgo para que el modelo tenga una estimación inicial independiente de las variables predictoras. Esta es un convenciòn comùn en la regresiòn lineal.

Existe dos tipos de sesgo:

- **Sesgo positivo:** Las predicciones del modelo son más altas que los valores reales.
- **Sesgo negativo:** Las predicciones son más bajas que los valores reales.

Un modelo ideal tiene un sesgo cercano a 0, lo que significa que las predicciones están cercanas a los valores reales. Sin embargo, en la práctica es normal que los modelos tengan cierto nivel de sesgo debido a las simplificaciones y suposiciones que se hacen del modelado.

Paso 4: Determinar los datos de los coeficientes Usando Álgebra Lineal

Para calcular los coeficientes (b0, b1, b2, b3, b4) de la regresión lineal, utilizamos álgebra lineal y la matriz de diseño (X) y el vector de resultados (Y).

La matriz de diseño (X) se forma concatenando un vector de unos (para el término de sesgo) y las columnas de las características (A, B, C, D). El vector de resultados (Y) contiene los precios de las viviendas.

El resultado será un vector de coeficientes (b0, b1, b2, b3, b4) que indicará la importancia de cada característica para predecir el precio de las viviendas. Los coeficientes más grandes indicarán una mayor importancia.

Matriz de Diseño X (coeficientes):* ***Vector de Resultados Y (precios):****

1	2	1	1	0
1	3	2	0	1
1	4	3	1	1
1	4	1	1	0
1	3	2	1	1
1	2	1	0	1
1	4	2	1	0

150
200
250
170
220
180
210

Ahora, para encontrar los coeficientes (B), utilizamos la fórmula (método de mínimos cuadrados ordinarios):

B = (X^T * X)^(-1) * X ^T *Y

- **X^T:** Transpuesta de la matriz de características X.
- **X:** Matriz de características, que incluye una columna de unos para el término constante y las característica predictoras.
- **(X^T * X) ^(-1):** Inversa de la matriz resultante de multiplicar la transpuesta de X por X.
- **X ^T * Y:** Producto de la transpuesta de X por el vector de respuesta Y.
- **B:** Coeficiente que minimizan la suma de los cuadrados de los residuos.

Paso 5: Calcular los Coeficientes (B)

Primero, calculamos `(X^T * X)` y `(X^T * Y)`.

Matriz `(X^T )`:

1	1	1	1	1	1	1
2	3	4	4	3	2	4
1	2	3	1	2	1	2
1	0	1	1	1	0	1
0	1	1	0	1	1	0

Matriz `(X^T * X)`:

1	1	1	1	1	1	1
2	3	4	4	3	2	4
1	2	3	1	2	1	2
1	0	1	1	1	0	1
0	1	1	0	1	1	0

*

1	2	1	1	0
1	3	2	0	1
1	4	3	1	1
1	4	1	1	0
1	3	2	1	1
1	2	1	0	1
1	4	2	1	0

=

7	22	12	5	4
22	74	40	17	12
12	40	24	9	8
5	17	9	5	2
4	12	8	2	4

FILA 1

(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)+(1*1)=7

(1*2)+(1*3)+(1*4)+(1*4)+(1*3)+(1*2)+(1*4)=22

(1*1)+(1*2)+(1*3)+(1*1)+(1*2)+(1*1)+(1*2)=12

(1*1)+(1*0)+(1*1)+(1*1)+(1*1)+(1*0)+(1*1)=5

(1*0)+(1*1)+(1*1)+(1*0)+(1*1)+(1*1)+(1*0)=4

FILA 2

(2*1)+(3*1)+(4*1)+(4*1)+(3*1)+(2*1)+(4*1)=22

(2*2)+(3*3)+(4*4)+(4*4)+(3*3)+(2*2)+(4*4)=74

(2*1)+(3*2)+(4*3)+(4*1)+(3*2)+(2*1)+(4*2)=40

(2*1)+(3*0)+(4*1)+(4*1)+(3*1)+(2*0)+(4*1)=17

(2*0)+(3*1)+(4*1)+(4*0)+(3*1)+(2*1)+(4*0)=12

FILA 3

(1*1)+(2*1)+(3*1)+(1*1)+(2*1)+(1*1)+(2*1)=12

(1*2)+(2*3)+(3*4)+(1*4)+(2*3)+(1*2)+(2*4)=40

(1*1)+(2*2)+(3*3)+(1*1)+(2*2)+(1*1)+(2*2)=24

(1*1)+(2*0)+(3*1)+(1*1)+(2*1)+(1*0)+(2*1)=9

(1*0)+(2*1)+(3*1)+(1*0)+(2*1)+(1*1)+(2*0)=8

FILA 4

(1*1)+(0*1)+(1*1)+(1*1)+(1*1)+(0*1)+(1*1)=5

(1*2)+(0*3)+(1*4)+(1*4)+(1*3)+(0*2)+(1*4)=17

(1*1)+(0*2)+(1*3)+(1*1)+(1*2)+(0*1)+(1*2)=9

(1*1)+(0*0)+(1*1)+(1*1)+(1*1)+(0*0)+(1*1)=5

(1*0)+(0*1)+(1*1)+(1*0)+(1*1)+(0*1)+(1*0)=2

FILA 5

(0*1)+(1*1)+(1*1)+(0*1)+(1*1)+(1*1)+(0*1)=4

(0*2)+(1*3)+(1*4)+(0*4)+(1*3)+(1*2)+(0*4)=12

(0*1)+(1*2)+(1*3)+(0*1)+(1*2)+(1*1)+(0*2)=8

(0*1)+(1*0)+(1*1)+(0*1)+(1*1)+(1*0)+(0*1)=2

(0*0)+(1*1)+(1*1)+(0*0)+(1*1)+(1*1)+(0*0)=4

*****Matriz `(X^T * Y)`:*****

1	1	1	1	1	1	1		150		1380
2	3	4	4	3	2	4	*	200	=	4440
1	2	3	1	2	1	2		250		2510
1	0	1	1	1	0	1		170		1000
0	1	1	0	1	1	0		220		850
								180		
								210		

FILA 1

(1*150)+(1*200)+(1*250)+(1*170)+(1*220)+(1*180)+(1*210)=1380

FILA 2

(2*150)+(3*200)+(4*250)+(4*170)+(3*220)+(2*180)+(4*210)=4440

FILA 3

(1*150)+(2*200)+(3*250)+(1*170)+(2*220)+(1*180)+(2*210)=2510

FILA 4

(1*150)+(0*200)+(1*250)+(1*170)+(1*220)+(0*180)+(1*210)=1000

FILA 5

(0*150)+(1*200)+(1*250)+(0*170)+(1*220)+(1*180)+(0*210)=850

****Paso 5: Calcular los Coeficientes (B)****

Primero, calculamos la inversa de la matriz `(X^T * X)`:

7	22	12	5	4	1	0	0	0	0
22	74	40	17	12	0	1	0	0	0
12	40	24	9	8	0	0	1	0	0
5	17	9	5	2	0	0	0	1	0
4	12	8	2	4	0	0	0	0	1

F1=F1/7

1	22/7	12/7	5/7	4/7	1/7	0	0	0	0
22	74	40	17	12	0	1	0	0	0
12	40	24	9	8	0	0	1	0	0
5	17	9	5	2	0	0	0	1	0
4	12	8	2	4	0	0	0	0	1

F2=22*F1-F2 F3=12*F1-F3 F4=5*F1-F4 F5=4*F1-F5

1	22/7	12/7	5/7	4/7	1/7	0	0	0	0
0	34/7	16/7	9/7	-4/7	-22/7	1	0	0	0
0	16/7	24/7	3/7	8/7	-12/7	0	1	0	0
0	9/7	3/7	10/7	-6/7	-5/7	0	0	1	0
0	-4/7	8/7	-6/7	12/7	-4/7	0	0	0	1

F2=F2/(34/7)

1	22/7	12/7	5/7	4/7	1/7	0	0	0	0
0	1	8/17	9/34	-2/17	-11/17	7/34	0	0	0

0	16/7	24/7	3/7	8/7	-12/7	0	1	0	0
0	9/7	3/7	10/7	-6/7	-5/7	0	0	1	0
0	-4/7	8/7	-6/7	12/7	-4/7	0	0	0	1

F1=-22/7*F2+F1 F3=-16/7*F2+F3 F4=-9/7*F2+F4 F5=-(-4/7*F2)+F5

1	0	4/17	-2/17	16/17	37/17	-11/17	0	0	0
0	1	8/17	9/34	-2/17	-11/17	7/34	0	0	0
0	0	40/17	-3/17	24/17	-4/17	-8/17	1	0	0
0	0	-3/17	37/34	-12/17	2/17	-9/34	0	1	0
0	0	24/17	-12/17	28/17	-16/17	2/17	0	0	1

F3=F3/(40/17)

1	0	4/17	-2/17	16/17	37/17	-11/17	0	0	0
0	1	8/17	9/34	-2/17	-11/17	7/34	0	0	0
0	0	1	-3/40	3/5	-1/10	-1/5	17/40	0	0
0	0	-3/17	37/34	-12/17	2/17	-9/34	0	1	0
0	0	24/17	-12/17	28/17	-16/17	2/17	0	0	1

F1=-4/17*F3+F1 F2=-8/17*F3+F2 F4=-(-3/17)*F3+F4 F5=-24/17*F3+F5

1	0	0	-1/10	4/5	11/5	-3/5	-1/10	0	0
0	1	0	3/10	-2/5	-3/5	3/10	-1/5	0	0
0	0	1	-3/40	3/5	-1/10	-1/5	17/40	0	0
0	0	0	43/40	-3/5	1/10	-3/10	3/40	1	0
0	0	0	-3/5	4/5	-4/5	2/5	-3/5	0	1

F4=F4/(43/40)

1	0	0	-1/10	4/5	11/5	-3/5	-1/10	0	0
0	1	0	3/10	-2/5	-3/5	3/10	-1/5	0	0
0	0	1	-3/40	3/5	-1/10	-1/5	17/40	0	0
0	0	0	1	-24/43	4/43	-12/43	3/43	40/43	0
0	0	0	-3/5	4/5	-4/5	2/5	-3/5	0	1

F1=-(-1/10)*F4+F1 F2=-3/10*F4+F2 F3=-(-3/40)*F4+F3 F5=-(-3/5)*F4+F5

1	0	0	0	32/43	95/43	-27/43	-4/43	4/43	0
0	1	0	0	-10/43	-27/43	33/86	-19/86	-12/43	0
0	0	1	0	24/43	-4/43	-19/86	37/86	3/43	0
0	0	0	1	-24/43	4/43	-12/43	3/43	40/43	0
0	0	0	0	20/43	-32/43	10/43	-24/43	24/43	1

F5=F5/(20/43)

1	0	0	0	32/43	95/43	-27/43	-4/43	4/43	0
0	1	0	0	-10/43	-27/43	33/86	-19/86	-12/43	0
0	0	1	0	24/43	-4/43	-19/86	37/86	3/43	0
0	0	0	1	-24/43	4/43	-12/43	3/43	40/43	0
0	0	0	0	1	-8/5	1/2	-6/5	6/5	1

F1=-32/43*F5+F1 F2=-(-10/43)*F5+F2 F3=-24/43*F5+F3 F4=-(-24/43)*F5+F4

1	0	0	0	0	17/5	-1	4/5	-4/5	-8/5
0	1	0	0	0	-1	1/2	-1/2	0	1/2
0	0	1	0	0	4/5	-1/2	11/10	-3/5	-6/5
0	0	0	1	0	-4/5	0	-3/5	8/5	6/5
0	0	0	0	1	-8/5	1/2	-6/5	6/5	43/20

La matriz inversa es:

17/5	-1	4/5	-4/5	-8/5
-1	1/2	-1/2	0	1/2
4/5	-1/2	11/10	-3/5	-6/5
-4/5	0	-3/5	8/5	6/5
-8/5	1/2	-6/5	6/5	43/20

Luego, multiplicamos `(X^T * X)^(-1)` por `X^T * Y` para obtener el vector de coeficientes B:

17/5	-1	4/5	-4/5	-8/5		1380		100		B0
-1	1/2	-1/2	0	1/2	*	4440	=	10	=>	B1
4/5	-1/2	11/10	-3/5	-6/5		2510		25		B2
-4/5	0	-3/5	8/5	6/5		1000		10		B3
-8/5	1/2	-6/5	6/5	43/20		850		27.5		B4

FILA 1

(17/5*1380)+(-1*4440)+(4/5*2510)+(-4/5*1000)+(-8/5*850)=100

FILA 2

(-1*1380)+(1/2*4440)+(-1/2*2510)+(0*1000)+(1/2*850)=10

FILA 3

(4/5*1380)+(-1/2*4440)+(11/10*2510)+(-3/5*1000)+(-6/5*850)=25

FILA 4

(-4/5*1380)+(0*4440)+(-3/5*2510)+(8/5*1000)+(6/5*850)=10

FILA 5

(-8/5*1380)+(1/2*4440)+(-6/5*2510)+(6/5*1000)+(43/20*850)=27.50

Entonces, los coeficientes de la regresión lineal son:

-b0 ≈ 100 (intercepto).

-b1 ≈ 10 (coeficiente A)

-b2 ≈ 25 (coeficiente B)

-b3 ≈ 10 (coeficiente C)

-b4 ≈ 27.5 (coeficiente D)

Los coeficientes más grandes (b4, b2) indican una mayor importancia en la predicción del precio de las viviendas. Por lo tanto, en función de la regresión lineal, las características de si se tiene estacionamiento (D) y la cantidad de baños (B), son las más importantes para predecir el precio de las viviendas en este conjunto de datos.

Importancia de los datos mas relevantes para la IA

- **Selección de Características:** En la IA se puede utilizar los resultados de la regresión lineal para las características más válidas en un grupo de datos. En el ejercicio, se encontró que el B y D son coeficientes grandes. Lo cual es válido por la selección de características que permite a la IA centrarse en lo más importante y descartar lo que no.
- **Predicciones más precisas:** Conociendo las características más importantes, la IA puede realizar predicciones más precisas. Al asignar un peso adecuado a cada característica en función de sus coeficientes, el modelo puede ajustarse mejor a los datos y producir predicciones más cercanas a la realidad.
- **Reducción de la Dimensionalidad:** En el contexto de la regresión lineal, es posible que algunas características tengan coeficientes cercanos a cero. La IA puede utilizar esta información para reducir la dimensionalidad del conjunto de datos, lo que ahorra recursos computacionales y acelera los cálculos sin perder precisión en las predicciones.

Conclusión

Abordar la maldición de la dimensionalidad en el aprendizaje automático mediante la aplicación de técnicas de reducción supervisada con regresión lineal ofrece una solución efectiva. Al emplear la combinación lineal, matriz transpuesta, reducción de matriz y matriz inversa, se logra optimizar la IA al reducir la complejidad de los datos, superando así los desafíos asociados con la alta dimensionalidad y mejorando la eficiencia de los modelos, evitando el sobreajuste y reduciendo los tiempos de entrenamiento prolongados.

Bibliografia

Delua, J. (12 de marzo de 2021). *ibm*. Obtenido de ibm: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Gonzalez, L. (30 de noviembre de 2018). *aprendeia*. Obtenido de aprendeia: https://aprendeia.com/algoritmo-regresion-lineal-simple-machine-learning/#:~:text=La%20Regresión%20Lineal%20puede%20ser,contribuyen%20a%20la%20variable%20dependiente.

Gonzalez, L. (9 de noviembre de 2018). *aprendeia*. Obtenido de aprendeia: https://aprendeia.com/bias-y-varianza-en-machine-learning/

javatpoint. (s.f.). *javatpoint*. Obtenido de javatpoint: https://www.javatpoint.com/linear-algebra-for-machine-learning

Rodríguez, E. (s.f.). *canalinnova*. Obtenido de canalinnova: https://canalinnova.com/la-maldicion-de-la-dimensionalidad-en-machine-learning-un-analisis/