Regresión Ridge

Santiago Laplagne

12 de junio de 2024

1. Introducción

En estas notas desarrollamos los conceptos teóricos de álgebra lineal detrás de regresión Ridge.

1.1. Regresión lineal y colinealidad

Decimos que un conjunto de datos presenta colinealidad si hay una relación de dependencia lineal entre las variables.

Ejemplo En esta base de exportaciones de Argentina por año, ¿que variables presentan dependencia lineal?

EXPO_AGROP	EXPO_PECUAR	EXPO_AGRIC	EXPO_CEREAL	EXPO_OLEAGI	EXPO
5675922.71	2220807.38	3450986.58	2579153.98	763762.94	10
5254582.56	2410845.43	2839752.13	1906466.84	794068.11	13
3685459.82	2277316.06	1403969.22	847878.88	245962.74	3
4350920.16	2755310.97	1592394.96	1080429.13	218880.92	29
3341704.72	1441522.79	1888866.22	1260219.64	294294.64	33
3394203.42	1621378.16	1758401.13	1170262.33	291286.07	29
3289516.91	2033833.42	1232810.26	763813.07	157240.83	31
3408794.66	1585888.12	1767175.1	1209749.33	238334.59	31
2999981	935528.07	1984893.91	1310043.44	198589.52	47

- EXPO_AGROP = EXPO_PECUAR + EXPO_AGRIC
- EXPO_AGRIC = EXPO_CEREAL + EXPO_OLEAGI + EXPO_OTAGRIC

1.2. Si hay colinealidad no hay unicidad

Cuando hay colinealidad, no podemos aplicar mínimos cuadrados directamente. El problema de mínimos cuadrados no tiene solución única.

Ejemplo: si queremos modelar el PBI utilizando las variables de exportación, estás fórmulas son equivalentes.

- PBI $\simeq$ EXPO_PECUAR + EXPO_AGRIC + ...

- PBI $\simeq$ EXPO_AGROP + ...
- PBI $\simeq$ EXPO_PECUAR + EXPO_CEREAL + EXPO_OLEAGI + EXPO_OTAGRIC + ...

1.3. Explosión de coeficientes

Peor aún, podemos obtener fórmulas con coeficientes enormes.

Estas fórmulas también son equivalentes:

- PBI $\simeq$ EXPO_AGROP $+ \cdots$
- PBI $\simeq 1001 \cdot$EXPO_AGROP$- 1000 \cdot$EXPO_PECUAR$- 1000 \cdot$EXPO_AGRIC$+ \cdots$

Fórmulas con coeficientes tan grandes (que se cancelan mutuamente) van a traer problemas numéricos y afectan severamente el modelo.

1.4. ¿Cómo tratamos la colinealidad?

Existen diversos métodos para lidiar con la colinealidad.

En este ejemplo, la opción más simple sería eliminar variables redundantes (por ejemplo, EXPO_AGROP y EXPO_OTAGRIC).

Existen también diversos métodos para detectar la colinealidad. Por ejemplo, podemos triangular la matriz por columnas.

1.5. ¿Y cómo tratamos la "casi colinealidad"?

Muchas veces la relación lineal no es exacta (por errores numéricos u otros motivos). Las variables pueden ser en teoría linealmente independientes pero en la práctica afectar igualmente el modelado.

En estos casos, podemos usar métodos para reducción de dimensionalidad (disminuir la cantidad de variables sin perder mucha información).

Alternativamente, podemos modificar el método de mínimos cuadrados para atacar la colinealidad.

1.6. Solución única y explosión de coeficientes

Consideramos estos dos conjuntos de datos. Queremos explicar y utilizando las variables x como predictoras (sin intercept).

Modelo: $y = \beta_1 x_1$

y	x_1
1	1.001
0	0.001
0	0.001

Modelo: $y = \beta_1 x_1 + \beta_2 x_2$

y	x_1	x_2
1	1.001	1.000
0	0.001	0.001
0	0.001	0.001

Preguntas

1. En el primer caso, ¿hay solución exacta? ¿Cuál esperan que sea la solución de mínimos cuadrados?
2. En el segundo caso, las variables x_1 y x_2 ¿son linealmente independientes?
3. ¿Hay solución exacta en este caso? ¿Cuáles son los coeficientes de la solución?

Respuestas

1. No hay solución exacta. La solución de mínimos cuadrados es $\beta_1 = 0.999998$.
2. Las variables x_1 y x_2 son linealmente independientes.
3. Este sistema tiene solución exacta: $y = 1000x_1 - 1000x_2$.

Si bien en el segundo modelo el error en entrenamiento es menor (porque la solución es exacta), un modelo con coeficientes tan grandes generalmente va a funcionar mal, va a tener menor capacidad predictiva.

Podemos pensar el segundo modelo como un caso de sobreajuste.

2. Mínimos cuadrados regularizados

Para evitar el problema del ejemplo anterior, queremos construir un modelo donde los coeficientes sean pequeños, o lo más chicos posibles.

Es decir, queremos dos cosas:

- Que los errores del modelo sean lo más chicos posibles.
- Que los coeficientes del modelo sean lo más chicos posibles

Podemos pensar la segunda condición como *buscar un modelo lo más simple posible*.

2.1. Solución: penalizamos a los modelos complicados

Como el problema de mínimos cuadrados en general tiene solución única, podemos minimizar una de las dos cosas, pero no las dos a la vez.

Penalidades. La solución común cuando queremos minimizar varias cosas a la vez es introducir penalidades.

En este caso, queremos agregar al problema de mínimos cuadrados una penalidad cuando los errores son grandes.

2.2. Función de pérdida

Función de pérdida. Queremos definir una función de pérdida del estilo

$$L = \text{error cuadrático medio del ajuste} + \alpha \cdot \text{ tamaño de los coeficientes}$$

para algún parámetro α apropiado de penalidad. α es el *peso* que le damos a la penalidad en el modelo.

Intuitivamente, entre dos modelos que ajustan razonablemente bien, elegimos el que tiene coeficientes más chicos.

2.3. Formulación matemática

El error cuadrático del modelo se calcula por la fórmula:

$$EC = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_m - \hat{y}_m)^2,$$

donde y_i son los valores reales e $\hat{y}_i$ son las predicciones del modelo.

Comparamos esta fórmula con la norma al cuadrado de un vector:

$$\|v\|^2 = \|(v_1, \ldots, v_m)\|^2 = v_1^2 + v_2^2 + \cdots + v_m^2$$

Observamos que

$$EC = \|(y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_m - \hat{y}_m)\|^2 = \|y - \hat{y}\|^2.$$

Por lo tanto: Minimizar el error cuadrático es equivalente a minimizar $\|y - \hat{y}\|^2$.

2.4. Formulación matemática

A la vez, podemos medir el tamaño de los coeficientes por la norma al cuadrado del vector de coeficientes:

$$\beta_1^2 + \beta_2^2 + \cdots + \beta_n^2 = \|(\beta_1, \beta_2, \ldots, \beta_n)\|^2$$

Si ahora juntamos las dos fórmulas, obtenemos una función de pérdida:

$$\begin{aligned} L &= \text{error cuadrático medio del modelo} + \alpha \cdot \text{ tamaños de los coeficientes} \\ &= \|y - \hat{y}\|^2 + \alpha \|\beta\|^2 \end{aligned}$$

Es decir, queremos que tanto la norma del vector de errores como la norma del vector de coeficientes sean pequeños.

Obtenemos así el método conocido como *regresión Ridge* (o mínimos cuadrados regularizados o regulización L_2 o regularización de Tychonov).

Es un método muy común y utilizado en diversas áreas, por eso los diversos nombres.

La función de pérdida es

$$L = \|y - \hat{y}\|_2^2 + \alpha \|\beta\|_2^2$$

para un parámetro α a determinar.

2.5. El milagro de los mínimos cuadrados regularizados

Ya mencionamos que el problema de mínimos cuadrados puede resolverse fácilmente utilizando álgebra lineal.

El problema de mínimos cuadrados regularizados parece mucho más complicado.

Milagrosamente, este problema también puede resolverse fácilmente utilizando álgebra lineal.

Dado un sistema de ecuaciones $X\beta = y$, para hallar el vector β que minimiza

$$L = \|y - \hat{y}\|_2^2 + \alpha\|\beta\|_2^2$$

resolvemos el sistema lineal

$$\left(X^\mathsf{T}X + \alpha I\right)\beta = X^\mathsf{T}y.$$

2.6. Parámetros e hiperparámetros

Parámetros. En el modelo lineal, una vez que fijamos una fórmula $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$, calculamos los valores óptimos de $\beta_0, \ldots, \beta_n$ entrenando el modelo en los datos de entrenamiento.

Los coeficientes $\beta_0, \ldots, \beta_n$ se denominan *parámetros* del modelo.

Hiperparámetros. En el modelo lineal regularizado, debemos fijar primero un valor de α y luego podemos calcular los valores óptimos de $\beta_0, \ldots, \beta_n$ entrenando el modelo en los datos de entrenamiento. En este caso α se denomina un *hiperparámetro* del modelo.

Los hiperparámetros no aprenden su valor de los datos, sino que debemos especificarlos "manualmente".

2.7. ¿Cómo elegimos un valor apropiado para el hiperparámetro α?

Una forma usual para elegir el valor del hiperparámetro es por validación cruzada en k pliegos.

Realizamos los siguientes pasos:

1. **Preparación de datos**: Dividimos los datos en conjunto de entrenamiento y conjunto de testeo.
2. **Selección de parámetros**: Definimos un vector de posibles valores para el hiperparámetro α.
3. **Ajuste del modelo**: Para cada valor de α, ajustamos un modelo Ridge utilizando los datos de entrenamiento y calculamos su rendimiento utilizando validación cruzada.

4. **Validación cruzada**: Los datos de entrenamiento se dividen en k pliegues (folds), y el modelo se entrena k veces, cada vez utilizando $k-1$ pliegues para el entrenamiento y el pliegue restante para la validación. Este proceso se repite k veces, de manera que cada pliegue se utiliza una vez como conjunto de validación. El error del modelo se promedia sobre las k iteraciones.

5. **Selección del mejor valor de** α: Seleccionamos el valor de α que minimiza el promedio de los errores en las k iteraciones.

6. **Entrenamiento final**: Una vez que seleccionamos el mejor valor de α, ajustamos un modelo Ridge final utilizando todos los datos de entrenamiento y este valor de α.

7. **Evaluación final**: Evaluamos el modelo final utilizando el conjunto de prueba reservado anteriormente para obtener una estimación imparcial de su rendimiento en datos no vistos.