Pregunta 2.

1. Construya un ejemplo a mano 3×4, i.e, 3 variables predictoras y una a predecir, en el que se verifique que da el mismo resultado si los β's se calculan usando la forma clásica...

Sea $X = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ y $Y = \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}$

Aplicando a cada columna de X Gram-Schmidt se tiene

$$u_1 = x_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$u_2 = x_2 - \frac{\langle u_1, x_2 \rangle}{\langle u_1, u_1 \rangle} u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \frac{\left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\rangle}{\left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

$$u_3 = x_3 - \frac{\langle u_1, x_3 \rangle}{\langle u_1, u_1 \rangle} u_1 - \frac{\langle u_2, x_3 \rangle u_2}{\langle u_2, u_2 \rangle} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\Rightarrow e_1 = \frac{u_1}{\|u_1\|} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad e_2 = \frac{u_2}{\|u_2\|} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad e_3 = \frac{u_3}{\|u_3\|} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Ahora $X \cdot \beta = \langle y, e_1 \rangle e_1 + \langle y, e_2 \rangle e_2 + \langle y, e_3 \rangle e_3$

$$= 4 e_1 + 3 e_2 + 5 e_3$$

$$= \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix} = y.$$

$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^t = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$

Con el método de regresión

$$\beta = \left((X^t X)^{-1} X^t y\right) = \left(\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^t \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)^{-1} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}$$

$$= \left(\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)^{-1} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ 7 \\ 2 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 7 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} \quad \therefore \beta = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

$$\Rightarrow y = X \cdot \beta = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix} /\!/$$

$\therefore$ Si se llega con ambos a la mismo y.

2 Pba. Sea $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$, $x \in \mathbb{R}^n$.

Sea $y = Pb \in \text{Rango}(A)$; P es la proyección ortogonal

y minimiza $\|r\|_2 = \|Ax - b\|_2$

Ahora sea $z \in \text{Rango}(A)$ con $z \neq y$.

$z-y$ es ortogonal a $b-y$ $\Rightarrow \|b-z\|_2^2 = \|b-y\|_2^2 + \|y-z\|_2^2$

$\Rightarrow \|b-z\|_2 > \|b-y\|_2 \; ; \forall z$.

Por lo tanto y minimiza $\|b-z\|_2$, entonces $Ax = Pb \in \text{Rango}(A)$

Pregunta 6: Demuestre que Ridge puede ser obtenido mediante Regresión Lineal usando una versión aumentada de la tabla de datos. Se aumenta X con p filas adicionales $\sqrt{\lambda}\, I$ y se aumenta y con p ceros:

$$\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} y \\ \vec{0} \end{bmatrix}$$

donde $\vec{0} = (0,0,\dots,0)^T$, osea $\tilde{X}$ es de tamaño $(n+p)\times m$ que $\tilde{y}$ es de tamaño $(n+p)\times 1$. Mediante ...

Sea X una tabla $n \times m$ y Y un vector.

$$\Rightarrow \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix}; \quad \tilde{y} = \begin{pmatrix} y \\ \vec{0} \end{pmatrix}$$

$\tilde{X}$ es de dimensión $(n+p)\times m$ y $\tilde{y}$ $(n+p)\times 1$.

$$\Rightarrow \tilde{X}^t = [X^t \ \sqrt{\lambda} I^t].$$

Ahora $\beta = (\tilde{X}^t \tilde{X})^{-1} \tilde{X}^t \cdot \tilde{y}$

$$= \left((X^t \ \sqrt{\lambda} I^t) \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix} \right)^{-1} (X^t \ \sqrt{\lambda} I) \begin{pmatrix} y \\ \vec{0} \end{pmatrix}.$$

$$= (X^t X + \lambda I)^{-1} X^t y + \sqrt{\lambda} I \cdot \vec{0}.$$

$$= (X^t X + \lambda I)^{-1} X^t y \quad //$$

Pregunta 7.

Regresión con modelo de p parámetros, ajustados por mínimos cuadrados. en un conjunto de datos de entrenamiento $(x_1, y_1), \ldots, (x_n, y_n)$ ...

· Sea $\hat{\beta}$ estimado de mínimos cuadrado.

· datos prueba $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)$.

Si $R_{tr}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \beta)^2$ y $R_{te}(\beta) = \frac{1}{m} \sum_{i=1}^{m} (\tilde{y}_i - \tilde{x}_i \beta)^2$.

Pruebe :

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})].$$

donde la esperanza es calculada en todo el espacio aleat...

La esperanza de $R_{te}(\beta)$ es igual a la esperanza de $(\tilde{y}_1 - \beta x_1)^2$, por lo tanto es independiente de m.

Tomando $m = n$ y sustituyendo en R_{te} $\hat{\beta}$ por β que minimiza la expresión, tendríamos que las medias de ambas son iguales.

Pregunta 8

a) Ejecutamos una regresión Ridge con parámetro λ en X. se obtiene el coeficiente a. Ahora incluimos una copia exacta $X^* = X$ y volvemos a calcular Ridge. Demuestre que ambos coef. son iguales y calcule su valor. Demuestre en general que si m copias de la variable X_j son incluidas, todos sus coef. son iguales.

Ridge selecciona β para minimizar

$$\|y - \beta X\|^2 + \lambda \|\beta\|^2$$

el mínimo está dado por

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

como X es de una sola variable, i.e, $p=1$, esto quiere decir que $X^T X$ es un escalar y su inversa también es un escalar. 1

$$\Rightarrow \beta = \frac{X^T y}{X^T X + \lambda I}$$

Cuando incluimos una copia exacta, el problema de optimización se vuelve

$$\|y - X\beta_1 - X\beta_2\|^2 + \lambda \|\beta_1\|^2 + \lambda \|\beta_2\|^2$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - x_i \beta_1 - x_i \beta_2)^2 + \lambda \beta_1^2 + \lambda \beta_2^2 \quad ; \text{ pues } \beta_1, \beta_2 \text{ son escalares}$$

Derivando

$$2 \sum_{i=1}^{n} (y_i - x_i \beta_1 - x_i \beta_2)(-x_i) + 2\lambda \beta_1 = 0$$

$$2 \sum_{i=1}^{n} (y_i - x_i \beta_1 - x_i \beta_2)(-x_i) + 2\lambda \beta_2 = 0$$

en una forma matricial esto es: y por regla de Cramer la solución es:

$$\begin{pmatrix} X^TX+\lambda & -X^TX \\ -X^TX & X^TX+\lambda \end{pmatrix} \left(\frac{1}{(X^TX+\lambda)(X^TX+\lambda)-(X^TX)^2} \right) = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\Rightarrow \quad \frac{X^Ty}{2X^TX+\lambda} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{nota: } \beta_1 = \beta_2$$

que era lo que buscabamos probar.

Ahora para m copias, el problema de optimización sería:

$$\left(y - X\sum_{j=1}^{m} \beta_j\right)^T \left(y - X\sum_{j=1}^{m} \beta_j\right) + \lambda \sum_{j=1}^{m} \beta_j^2$$

derivando e igualando a cero tenemos:

$$-X^T\left(y - X\sum_{j=1}^{m} \beta_j\right) - \left(y - X\sum_{j=1}^{m} \beta_j\right)X + 2\lambda\beta_K = 0; \forall K \in \{1,\dots,m\}$$

Lo cual podemos escribir como

$$X^TX\sum_{j=1}^{m} \beta_j + \lambda\beta_K = X^Ty.$$

Tomando otro $\beta_K \in \{\beta_1,\dots,\beta_m\}$ tendríamos las mismas ecuaciones.

$$\Rightarrow \quad mX^TX\beta_K + \lambda\beta_K = X^Ty$$

$$\Rightarrow \quad \beta_K = \frac{X^Ty}{mX^TX+\lambda} \quad ; \forall K \in \{1,\dots,m\} \;//$$