Devyani Mandia

PAGE No.
DATE / /

3. Variable Selection.

3.1 S → fixed integer → most relevant variables.
$X_{\cdot j}$ → all the entries of j^{th} column
Correlation $c_j = Cor(X_{\cdot j}, Y)$
S variables such that ⇒ ① Correlations are largest in absolute value; $|\hat{S}| = S$ & ② $|c_j| > |c_j|$ for any $j \in \hat{S}$, $j \notin \hat{S}$

⇒ Correlation → is a statistical measure that expresses till what extent two variables are linearly related. Or degree to which a pair of variables are linearly related.
But selecting a variable simply because its correlation with target is higher does not justify that they are actually good variables for our model.
For example, variables like house price as Y can be dependent on features such as house area, location, etc. but does not have any direct relation with the car parked in the garage. However when we calculate the correlation it may show otherwise, so it doesn't appear to be a good procedure for variable selection.
A spurious relationship occurs in such variables that shows association but not causally related, may occur due to either coincidence or presence of a 3rd unseen factor.
It is important that we consider advanced techniques for variable selection like lasso, as it has model selection consistency for highly correlated data.

3.2 Let $u \rightarrow p$-dimensional vector & $\lambda \in [0, \infty)$
$\underset{v \in R^p}{\operatorname{argmin}} \|u - v\|_2^2 + \lambda \|v\|_2^2$ $u \& \lambda$.

ds This function is similar to the ridge regression loss function.

Solution using multivariate calculus is as follows-
$v^* = (I_p + \lambda I_p)^{-1} u$ I_p = Identity matrix

Since $(I_p + \lambda I_p)^T$ is always invertible for $\lambda \in [0, \infty)$ which shows that the solution is a non-zero positive solution always exist.

If $v \neq 0$ then $(I + \lambda I)^{-1} v \neq 0$, since $v \neq 0$
So for any $j \in \mathbb{R}^p$ $v_j \neq 0$.

4.2 We use beta refit instead of beta lasso.
As lasso is basically a technique for variable selection and does not have a closed form. Our goal using beta_lasso is to only select variables which would truly impact the output & push the other variables (or suppress) to achieve 0 by pushing their β's or co-efficients to 0.

On the other hand beta_refit is using ridge regression formula $\hat{\beta}_{ridge} = (X^T X + \lambda I_p)^{-1} X^T Y$. Hence we have a closed form & we can use these coefficients β to actually estimate Y & get the test error using mean squared error.

5.1 It is an incorrect assumption that we will always find $X\beta = Y$ as the error term defined for this expression, i.e. $\|X\beta - Y\|_2^2$ is not actually always 0, a lot of times it is tending to 0 but it is our end goal to minimize this error but only a few instances of data has achieved the error as 0 but in normal data scenarios we try to minimize the term as much as possible so that $\|X\beta - Y\|_2^2 \to 0$ but achieving 0 may be a far fetched assumption even though the parameter p is 1. As usually there are instances of unaccounted noise in the data.

Hence in summary,
A p-dimensional vector β, $\beta \in \mathbb{R}^p$ such that

PAGE No.
DATE / /

(5.1) $X\beta = y$ & $p < n$ does not necessarily have a solution to satisfy these set of equations perfectly.

(5.2) $X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \in \mathbb{R}^{3\times 2}$ matrix

$$X^TX = X_{1\cdot}X_{1\cdot}^T + X_{2\cdot}X_{2\cdot}^T + X_{3\cdot}X_{3\cdot}^T$$

$(AB)^T = B^TA^T \quad \therefore (X_{1\cdot}X_{1\cdot}^T)^T = (X_{1\cdot}^T X_{1\cdot})^T$

Similarly $X_{2\cdot}X_{2\cdot}^T = (X_{2\cdot}^T X_{2\cdot})^T$
& $X_{3\cdot}X_{3\cdot}^T = (X_{3\cdot}^T X_{3\cdot})^T$

Now $(X_{1\cdot}^T X_{1\cdot})^T = \left[\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}\begin{pmatrix} x_{11} & x_{12} \end{pmatrix}\right]^T = \begin{pmatrix} x_{11}^2 & x_{11}x_{12} \\ x_{11}x_{12} & x_{12}^2 \end{pmatrix}^T = \begin{pmatrix} x_{11}^2 & x_{11}x_{12} \\ x_{11}x_{12} & x_{12}^2 \end{pmatrix}$ → (1)

$(X_{2\cdot}^T X_{2\cdot})^T = \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}\begin{pmatrix} x_{21} & x_{22} \end{pmatrix} = \begin{pmatrix} x_{21}^2 & x_{21}x_{22} \\ x_{21}x_{22} & x_{22}^2 \end{pmatrix}^T = \begin{pmatrix} x_{21}^2 & x_{21}x_{22} \\ x_{21}x_{22} & x_{22}^2 \end{pmatrix}$ → (2)

& $(X_{3\cdot}^T X_{3\cdot})^T = \begin{pmatrix} x_{31} \\ x_{32} \end{pmatrix}\begin{pmatrix} x_{31} & x_{32} \end{pmatrix} = \begin{pmatrix} x_{31}^2 & x_{31}x_{32} \\ x_{31}x_{32} & x_{32}^2 \end{pmatrix}$ → (3)

Adding (1)+(2)+(3) $\Rightarrow \begin{pmatrix} x_{11}^2+x_{21}^2+x_{31}^2 & x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32} \\ x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32} & x_{12}^2+x_{22}^2+x_{32}^2 \end{pmatrix}$ → (4)

$X^TX = \begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{pmatrix}\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \begin{pmatrix} x_{11}^2+x_{21}^2+x_{31}^2 & x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32} \\ x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32} & x_{12}^2+x_{22}^2+x_{32}^2 \end{pmatrix}$ → (5)

So from (4) & (5) $X^TX = (X_{1\cdot}^T X_{1\cdot})^T + (X_{2\cdot}^T X_{2\cdot})^T + (X_{3\cdot}^T X_{3\cdot})^T$
$= X_{1\cdot}X_{1\cdot}^T + X_{2\cdot}X_{2\cdot}^T + X_{3\cdot}X_{3\cdot}^T$ Hence proved.

PAGE No.
DATE / /

(Q3) $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{21} & a_{23} \end{pmatrix}$ & $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

$$BA = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} + a_{21}b_{12} & a_{12}b_{11} + a_{22}b_{12} & a_{13}b_{11} + a_{23}b_{12} \\ a_{11}b_{21} + a_{21}b_{22} & a_{12}b_{21} + a_{22}b_{22} & a_{13}b_{21} + a_{23}b_{22} \\ a_{11}b_{31} + a_{21}b_{32} & a_{12}b_{31} + a_{22}b_{32} & a_{13}b_{31} + a_{23}b_{32} \end{pmatrix}$$

$$tr(AB) = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} + a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32}$$

$$tr(BA) = a_{11}b_{11} + a_{21}b_{12} + a_{12}b_{21} + a_{22}b_{22} + a_{13}b_{31} + a_{23}b_{32}$$

Therefore, $tr(AB) = tr(BA)$.

Q.4 Let $\tilde{\beta} = X^T(XX^T)^{-1}y$

To prove: $\tilde{\beta}$ → solution for $\min_{\beta \in \mathbb{R}^p} \|\beta\|_2^2$

such that $X\beta = y$, $X \in \mathbb{R}^{n \times p}$ where $p \geq n$.

& XX^T → invertible.

Now, $\alpha \in \mathbb{R}^p$ where $X\alpha = 0$, we need $\alpha^T\tilde{\beta} = 0$.

$$\alpha^T\tilde{\beta} = \alpha^T X^T(XX^T)^{-1}y$$
$$= (X\alpha)^T(XX^T)^{-1}y$$
$$= 0 * (XX^T)^{-1}y \quad \because X\alpha = 0 \text{ (given)}$$

PAGE No.
DATE / /

$\tilde{\beta} \Rightarrow$ orthogonal to any vector that satisfies $Xa = 0$
& if can be expressed as linear combination of the rows of X

$XX^T \rightarrow$ invertible so $\tilde{\beta}$ can be expressed as a linear combination of rows of X

$\tilde{\beta}$ is the least square estimate to $X\beta = Y$ & is the solution of $\min_{\beta \in \mathbb{R}^p} \|\beta\|_2^2$ such that $X\beta = Y$

$\Rightarrow$ Hence proved