Lec 11 Math 390 3/14/19

$SST = SSR + SSE$

$\uparrow$

fixed! only a function of y and not X!

$\Rightarrow SSR \uparrow \Rightarrow SSE \downarrow \Rightarrow R^2 \uparrow$, $RMSE \downarrow$

$SSR \downarrow \Rightarrow SSE \uparrow \Rightarrow R^2 \downarrow$, $RMSE \uparrow$

$$\hat{\vec{y}} = H\vec{y} = QQ^T\vec{y} = \sum_{j=0}^{p} proj_{\vec{q}_j}(\vec{y})$$

Since each $\vec{q}_j$ is orthogonal $\Rightarrow$ $proj_{\vec{q}_j}(\vec{a})$ are all orthogonal since $\exists c\ \forall \vec{a}\ldots\ proj_{\vec{q}_j}(\vec{a}) = c\vec{q}_j$

$\Rightarrow$ by Pythag. Thm. in $p+1$ dimensions...

$$||\hat{\vec{y}}||^2 = \sum_{j=0}^{p} ||proj_{\vec{q}_j}(\vec{y})||^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2 - \sum_{i=1}^{n} 2\hat{y}_i\bar{y} + \sum_{i=1}^{n}\bar{y}^2$$

$$= ||\hat{\vec{y}}||^2 - 2\bar{y}\sum_{i=1}^{n}\hat{y}_i + n\bar{y}^2 = ||\hat{\vec{y}}||^2 - 2n\bar{y}^2 + n\bar{y}^2 = ||\hat{\vec{y}}||^2 - n\bar{y}^2$$

$$\Sigma\hat{y} = \hat{y}^T\vec{1}_n = (H\vec{y})^T\vec{1}_n = \vec{y}^T H^T\vec{1}_n = \vec{y}^T \underbrace{H\vec{1}_n}_{\vec{1}_n \in colsp[X]} = \vec{y}^T\vec{1}_n = \Sigma y_i = n\bar{y}$$

$$= \sum_{j=0}^{p} ||proj_{\vec{q}_j}(\vec{y})||^2 - n\bar{y}^2$$

What happens if we add a new feature $\vec{x}_{new}$ to X

$$X_{new} = [X \mid \vec{x}_{new}]$$ s.t. X_{new} is full rank w/ rank $= (p+1)+1$

What happens to SSR?

$$SSR_{new} = \sum_{j=0}^{p} \|proj_{e_j}(\vec{y})\|^2 + \underbrace{\|proj_{e_{new}}(\vec{y})\|^2}_{> 0} - n\bar{y}^2$$

new predictor

> 0 not $= 0$ otherwise X_{new} would not be full rank.

$$\Rightarrow SSR_{new} = SSR + \|proj_{e_{new}}(\vec{y})\|^2 > SSR$$

$\Rightarrow R^2_{new} > R^2$, $RMSE_{new} < RMSE$

We can increase R^2 by just adding a vector to X. Even a ~~totally~~ made up vector!! Demo..

What if we keep going.. keep adding columns until $p+1 = n$

X is square $n \times n$ and full rank $\Rightarrow X$ is now invertible

$$H = X(X^T X)^{-1} X^T = X X^{-1} (X^T)^{-1} X^T = I_n \quad \text{why?}$$

$(AB)^{-1} = B^{-1} A^{-1}$

Is I_n a proj matrix? Yes... onto $\mathbb{R}^n$ the whole space.

Two properties

$I^T = I$ ✓ eigenvalues of I? $\lambda_1 = \ldots = \lambda_n = 1$.

$I \cdot I = I$ ✓ eigenvectors of I? Any basis for $\mathbb{R}^n$ will do...

$\hat{y} = Hy = Iy = y$! $\hat{y} = y \Rightarrow e = y - \hat{y} = \vec{0}_n$!

$\Rightarrow SSE = 0 \Leftrightarrow RMSE = 0 \Leftrightarrow R^2 = 100\%$. Perfect fit!

All you need to do is fill up X with garbage random columns. Can this be real? NO!

It is called "in-sample" overfitting. (In-sample means within $\mathbb{D}$). It is a core concept in this class.

How did this happen? $\vec{x}_{\cdot p+2}, \vec{x}_{\cdot p+3}, \ldots \vec{x}_{\cdot n}$ are all reducing SSE by locating chance correlations on different dimensions of the residual space.

These vectors are only chance-correlated with y as they are not truly related to the z's in $t(z)$. Thus, this chance correlation must disappear!

Why is this bad? Because this class is focused on using the model g to make predictions on new data, $\hat{y}_* = g(\vec{x}_*)$. Overfitting is bad for future predictions. AKA out-of-sample (oos)

Bad for "generalization error" → how accurate is the model in the future?