241218 Lasso / Ridge regularization

L^P norm(=P-norm) : Method which is used to gauge magnitude or distance of vector in LP space (=Lebesgue space)

- L^P space(=Lebesgue space) : function space where is used to gauge the size of function. Especially, this space is defined as the set of functions whose absolute value raised to the power of p $(=f^P)$ is integrable. (=함수의 크기를 정의) tool,

$$L^P(\Omega)=\left\{f: \int_\Omega |f(x)|^P dx < \inf.\right\},\ s.t\ 1\le P<\infty$$

if $P=1 \Rightarrow$ use L^1 norm
if $P=2 \Rightarrow$ use L^2 norm...

$\Rightarrow L^P$ norm(=P norm) $= \|X\|_P = \left(\sum_i |x_i|^P\right)^{1/P}$, s.t $1\le P<\infty$ (=함수의 크기를 측정)

Lasso(=L1) : Least Absolute Shirnkage and Selection Operator $\Rightarrow$ Passive operator

↳ lasso is seemed select a few features which are relatively important.

↳ Lasso tries to minimize absolute value of weight value $(=opt(\sum_i |w_i|))$

loss function in Lasso regularization = MSE + $\lambda|w|$ $=L_1$

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \frac{\lambda}{2n}\sum_{i=1}^{n}|W_i|,$$

$$\frac{\partial L_1}{\partial w} = \frac{\partial}{\partial w}(wX+b-Y)^2 + \lambda\frac{|w|}{\partial w} = \begin{cases} 2X(wx+b-Y)+\lambda & s.t.\ w\ge 0 \\ 2X(wx+b-Y)-\lambda & s.t.\ w<0 \end{cases}$$

$\therefore$ let $A=2X(wx+b-Y)$, then $w^* = \begin{cases} w+\eta\cdot A-\lambda & s.t.\ w\ge 0 \\ w+\eta\cdot A+\lambda & s.t.\ w<0 \end{cases}$ $\Rightarrow$ Closer to Zero, and some values are even equal to zero.

↳ Effect of regulation

$\because$ every value are subtracted by λ equally. (=λ 계속 빼버면 언젠간 몇몇 w들은 0이 됨)

$\Rightarrow$ If assuming that all weight values have same scale, small weight values which are less importent will be deleted. (장르나 감독은 남기고 개봉연도나 상영시간은 없애나감)

Dif.

Ridge(=L2) loss function = MSE + λw^2 $=L_2$ $\Rightarrow$ Active operator

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 + \frac{\lambda}{2n}\sum_{i=1}^{n}W_i^2,$$

$$\frac{\partial L_2}{\partial w} = \frac{\partial}{\partial w}(wx+b-y)^2 + \lambda\frac{w^2}{\partial w} = 2X(wx+b-y) + 2\lambda w.$$

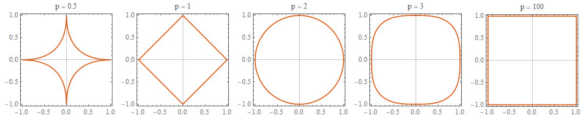
$\therefore$ let $A=2X(wx+b-y)$, then $w^* = w-\eta\cdot(A+2\lambda w)$ $\Rightarrow$ It operates similarly with L_1, but it regularize W^* more rationally.

Effect of regulation

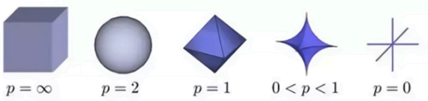
if W is big relatively, L_2 will subtract with bigger value, but if W is small, L_2 will subtract with smaller value. $\Rightarrow$ Sensitively activating to Outlier than L_1

cf) root is eliminated for enhancing efficiency of calculations.

⊗ Distribution of P-norm in 2-dimension vector space



⊗ Distribution of P-norm in 3-dimension vector space



Ref. https://seongyun-dev.tistory.com/52

+) Why do we want to make weight values close to zero?

= How large weights are related to overfitting)

If weights are small, the features have a small effect on prediction. the model fits the general trend of the data.

Else if weights are large, the model becomes highly sensitive to small changes to small changes in specific features.

In linear regression model, the prediction is:

$\hat{y} = \sum_i W_i x_i + b$, If W_i has large value, a small change in X_i will cause a large change in $\hat{y}$.