∴ Ridge regression is used to introduce bias to the data in order to generalize the data and increase bias.

This is useful if you dont have much training data.

Lasso Regression:
(L1 regularization / L1 Norm)

It is used to reduce the features. It helps in feature selection.

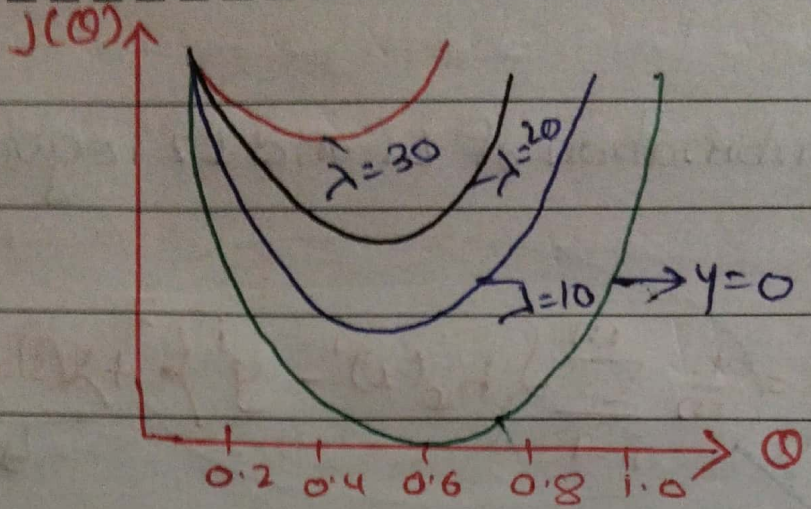
cost function

$$\text{cost function} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^i - y^{(i)}\right)^2 + \lambda \sum_{i=1}^{m} |\text{slope}|$$

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /



$$h_\Theta(x) = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \Theta_3 x_3$$

$$= \Theta_0 + 0.54x + 0.23 x_2 + \underline{0.1 x_3}$$

↳ least correlated

If data has outliers → use ridge regression

Lasso = least absolute shrinkage and selection operator regression

Lasso regression tends to eliminate the weights of the leadst important features by setting their weights to zero.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Elastic Net:

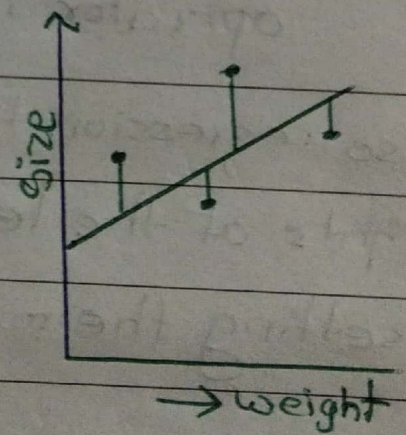
→ It is the combination of L1 and L2 regularization.

$$\text{Cost Function} = \frac{1}{m}\sum_{i=1}^{m}\left\{h_\theta(x)^i - y^i\right\}^2 + \underbrace{\lambda(\text{slope})^2}_{\hookrightarrow L2} + \underbrace{\lambda|\text{slope}|}_{\hookrightarrow L1}$$

Can be changed to MAE, RMSE, MSE

Note taken from John Stairmer

We find the line that results in the minimum sum of squared residuals.



∴ We end up with the eqⁿs of line

Say, Size = 0.95 + 0.75 × weight

0.75 → slope

Mo Tu We Th Fr Sa Su

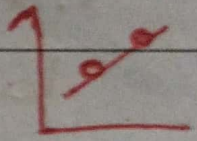
Memo No. _____

Date / /

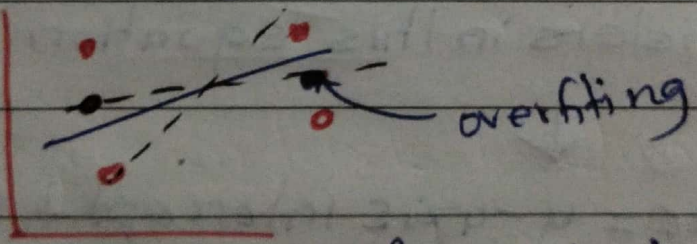
When we have a lot of measurements, we can be fairly confident that least square line accurately reflects the relationship bet^n size and weight.

But what if we only have two measurements?

→ we fit new line, since the new line overlaps the two data points, the minimum sum of squared residuals = 0.



∴ New line eq^n's: -



Sum of the squared residuals for testing data is large which means the new line has high variance.

∴ In ML, new line (blue) is overfit to training data.

Bias → amount that a model's prediction differs from the target value compare to training data.

Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /

∴ The main idea behind Ridge Regression is to find a new line that doesn't fit the training data as well.

In other words, we introduce a small amount of bias into how the new line is fit to the data but in return for that small amount of bias, we get a significant drop in variance.

∴ Ridge regression can provide better long term prediction.

When least square determines values for the parameters in this equation

↓

~~Size = y-axis intercept + slope * weight~~

⇓

It minimizes the sum of the squared residuals.

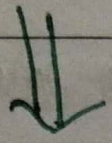
Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

In contrast,

When ridge regression determines values for the parameters in this equation



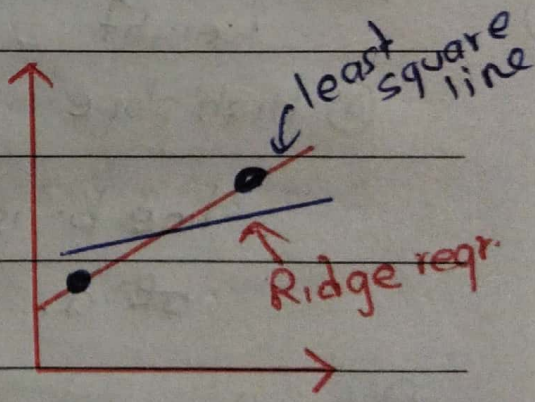
Size = y-axis intercept + slope * weight

It minimizes the sum of squared residuals

$+ \lambda * \text{the slope}^2$

lambda (λ) determines how severe that penalty is

this part adds a penalty to the traditional least squared method.



for red line:

$\lambda = 1$

$0 + \lambda(1.3)^2 = 0 + 1 \times 1.3^2$

$= 1.69$

for blue line

$\lambda = 1$

$(0.3)^2 + (0.1)^2 + \lambda(0.8)^2$

$= 0.74$

Mo Tu We Th Fr Sa Su

Memo No. _____

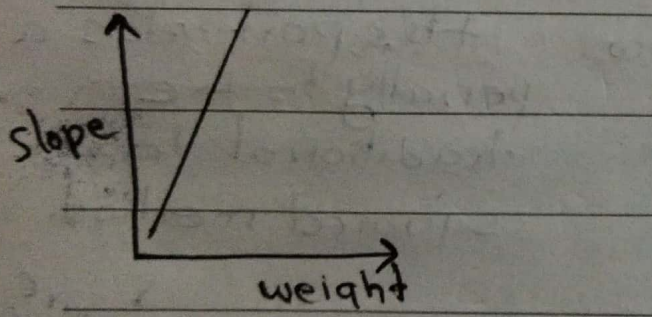
Date / /

Ridge Regression line:-

Red = 1.69, blue = 0.74

Thus, if we wanted to minimize the sum of the squared residuals plus the Ridge Regression penalty, we would choose the Ridge Regression line over the least square line.

→ Due to penalty in Ridge regression, small amount of bias, has less variance.



(a) High slope (b) low slope

· line b· is less sensitive to weight than of a.

Ridge Regression (RR):

Sum of squared residual + λ(slope)^2

→ λ can be any value from 0 to positive

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Infinity.

When $\lambda = 0$,

Ridge regression line = least square line

when $\lambda = 1$,

RR ended up with a smaller slope than the least square line.

and for larger λ,

the slope of RR gets asymptotically close to 0.

So, the larger λ gets, our prediction for size becomes less and less sensitive to weight.

So, how do we decide what value to give λ?

→ We just try a bunch of values for λ and use cross validation, typically 10 fold cross validation to determine which one results in the lowest variance.

"Uptill now RR was for continuous variable but RR works for discrete variable as well.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Ridge Regression can also be applied to logistic regression.

= the sum of likelihoods + $\lambda(\text{slope})^2$

Note:

When applied to logistic Regression, Ridge Regression optimizes the sum of the likelihoods instead of the squared residuals because logistic regression is solved using maximum likelihood.

"Ridge Regression helps reduce variance by shrinkage parameters and making our predictions less sensitive to them.

Summary:

When sample sizes are relatively small, RR can improve predictions made from new data (i.e reduce variance) by making the predictions less sensitive to the training data.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

RR pently itself by λ times the sum of all squared parameter, except for the y-intercept and λ is determined using cross validation.

Lasso Regression (L1):

Ridge regression penalty = $\lambda * (slope)^2$

Lasso Regression = $\lambda * |slope|$

Big difference between Ridge and Lasso Regression is that Ridge Regression can only shrink the slope asymptotically close to 0 while lasso regression can shrink the slope all the way to 0.

LR can exclude useless variables from eq's, better than RR at reducing the variance in models that contain a lot of useless variable.

Memo No. _____

Mo Tu We Th Fr Sa Su Date / /

Elastic Net Regression:

Elastic Net Regression starts with least squares then combines the lasso Regression penalty with the Ridge Regression penalty.

Sum of the squared residual

+

$\lambda_1 \times |\text{variable}| + \ldots |\text{variable}|$

+

$\lambda_2 * (\text{variable}_1)^2 + \ldots + (\text{variable}_n)^2$

Note: LR and RR penalty get their own λs.

The hybrid elastic net regression is especially good at dealing with situation when there are correlations between parameters.

By combining LR and RR,

→ Elastic Net Regression groups and shrink the parameters associated with the correlated variables and leaves them in eqn or removes them all at once.

Mo Tu We Th Fr Sa Su

Memo No. _____

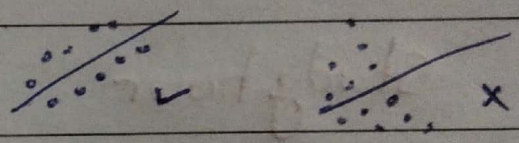
Date / /

Assumptions of linear Regression:-

① Linearity:-

→ linear relationship between Y and each X.

② Homoscedasticity:



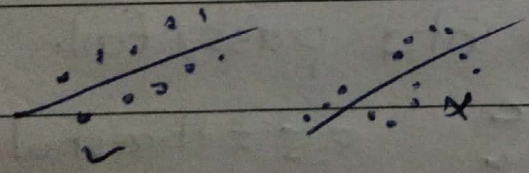
→ Equal variance.

③ Multivariate Normality

→ Normality of error distribution

④ Indepence.

→ of observation. includes no auto-correlation.



⑤ lack of multicollinearity:

Predictors are not correlated with each other.

$X_1 + X_2$ ✓ $X_1 \sim X_2$ ✗