Introduction

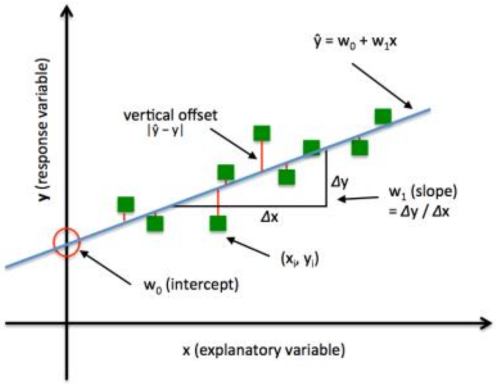
29 September 2022 06:41

Before we get directly into "Regularised Linear Regression", we could recall what do regression co-efficient tell us?, concept of overfitting, Bias and variance trade-off.

Regression Co-efficient:

- Recall from the normal linear regression
- The equation,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$



- We estimate the co-efficient using the least square criterion, which minimize the residual of sum of squares(RSS)
- In Graphically,
  - we're fitting the blue line to our data (the green points) that minimizes the sum of the distances between the points and the blue line (sum of the red lines) as shown above.
- Mathematically it is denoted as, (Whenever we want to bestfit)

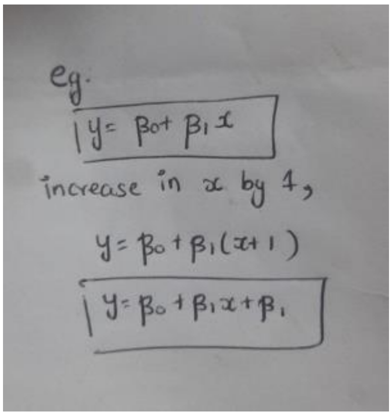
RSS = sum of (Actual - Predicted)^2

$$RSS = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \right)^2$$

- **n** is the total number of observations (data).
- **yi** is the **actual output value** of the observation (data).
- **p** is the total number of features.
- **βj** is a model's coefficient.
- **xij** is the ith observation, jth feature's value.
- **β0+∑βjxij** is the **predicted output** of each observation.

How increase in the independent variable increase the dependent variable?

- In regression with multiple independent variables, the coefficient tells you how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant.
- Below equation illustrate how it changes.



eg.

$\boxed{y = \beta_0 + \beta_1 x}$

increase in x by 1,

$y = \beta_0 + \beta_1(x+1)$

$\boxed{y = \beta_0 + \beta_1 x + \beta_1}$

- Hence the co-efficient is called as "Weight assigned to x"

OVERFITTING and Regularisation:

- when we built a model that is too complex that it matches the training data "too closely" or we can say that the model has started to learn not only the signal, but also the noise in the data.
- Resulting to,
  - Model do well on the training data
  - But, Failed to generalize the efficient result with unknown or data we have not seen before.
- Note:
  - Overfitting is due to High variance.
  - High variance is due to the high degree curve equation(or simply say, complexity)
  - Complexity by high number of features
  - Complexity by when the value of coefficient is high

Example:

X1	X2	X3	Y
0.1	1	10000	
0.2	2	20000	
0.3	3	30000	

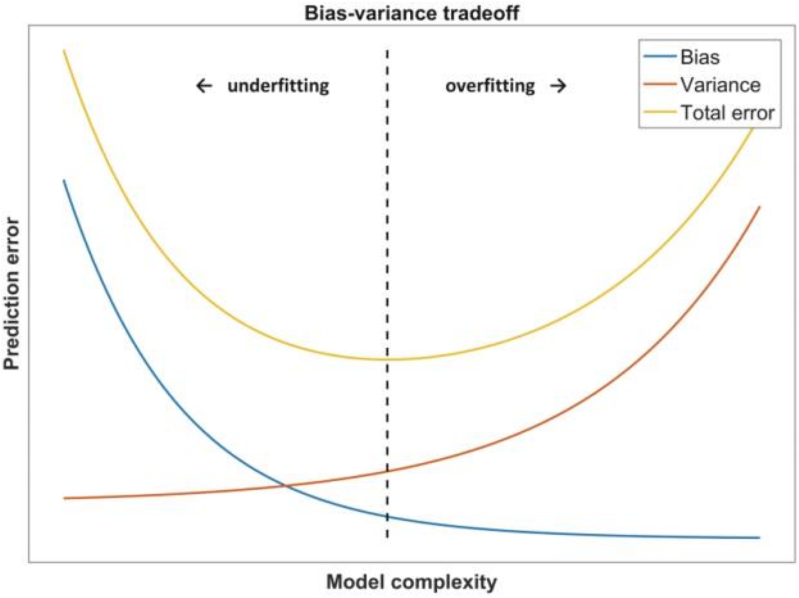
- In this example, we can conclude that, co-efficient of the x3 > x2 > x1.
- i.e. **β3>**β2>β1

- What if there an outlier in features?
  - While learning, equation not only learn the pattern,
  - But also learn the noise by the algorithm.
  - Looking at the above example, weight of co-efficient of the x3 will always be high.
- Hence, **Apart from removing the outliers how shall we reduce ?**

  Solution: "**Regularizing the variables**"
- This Regularization of the linear regression can be done using 3 ways
  1. Lasso regression
  2. Ridge regression
  3. Elastic net regression

BIAS & VARIANCE TRADE-OFF and Regularisation:

- Error in prediction models can be decomposed into the two sub-components
  - Error due to "bias"
  - Error due to "variance"
- Understanding these two will help to diagnose the model results and avoid the under/overfitting.



- **Bias**
  - The blue line, measures how far off in general our models' predictions are from the correct value.
  - Thus as our model gets more and more complex we will become more and more accurate about our predictions (Error steadily decreases).
- **Variance**
  - The red line, measures how different can our model be from one to another, as we're looking at different possible data sets.
  - If the estimated model will vary dramatically from one data set to the other, then we will have very erratic predictions, because our prediction will be extremely sensitive to what data set we obtain.
  - As the complexity of our model rises, variance becomes our primary concern.

Conclusion:

- When creating a model, our **goal** is to locate the **optimum model complexity**.
- If our model complexity exceeds this sweet spot, we are in effect overfitting our model; while if our complexity falls short of the sweet spot, we are underfitting the model.
- With all of that in mind, the **notion of regularization is simply a useful technique to use** when we think our model is too complex (models that have low bias, but high variance).
- It is a method for "constraining" or "regularizing" the size of the coefficients ("shrinking" them towards zero).
- The specific regularization techniques we'll be discussing are **Ridge Regression** and **Lasso Regression.**

Lasso and Ridge Regression

04 October 2022 23:32

- Regularized linear regression models are very similar to least squares, except that the coefficients are estimated by minimizing a slightly different objective function. we minimize the sum of RSS and a "penalty term" that penalizes coefficient size

  Linear Regression = RSS
  Lasso Regression (L1) minimizes **= $RSS + \lambda \sum|\beta j|$**
  Ridge Regression (L2) minimizes **= $RSS + \lambda \sum (\beta j)^2$**
- Lasso stands for **least absolute shrinkage and selection operator**
- **λ is a tuning parameter**
  - Which seeks to finding the balance between the fit of the model to the data and the magnitude of the model's coefficients
  - A **tiny λ** imposes no penalty on the coefficient size, and is equivalent to a normal linear regression.
  - **Increasing λ** penalizes the coefficients and thus shrinks them towards zero.
- RSS - **Cost function**
- $+\lambda \sum|\beta j|$ or $+\lambda \sum (\beta j)^2$ - **Penalty term**

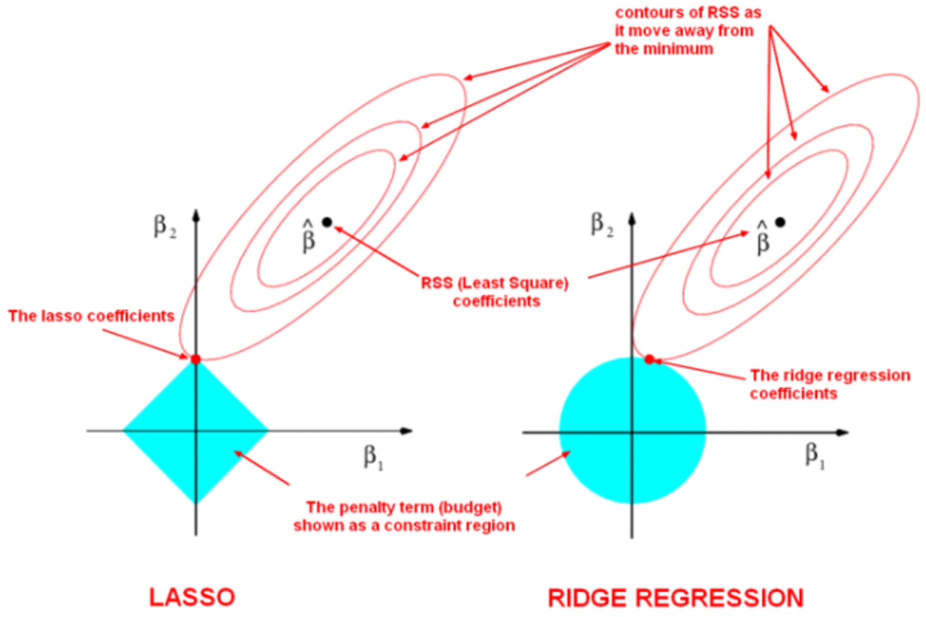
Hence, we can simply say that we are trying to balance two things to measure the model's total quality.

1. The RSS, measures how well the model is going to fit the data,
2. The magnitude of the coefficients, which can be problematic if they become too big.

- **Lasso regression** shrinks coefficients all the way to zero, thus removing them from the model.
- **Ridge regression** shrinks coefficients toward zero, but they rarely reach zero.

Visualizing Regularization

04 October 2022 23:51


contours of RSS as it move away from the minimum
$\hat{\beta}$
β_2
β_1
RSS (Least Square) coefficients
The lasso coefficients
The ridge regression coefficients
The penalty term (budget) shown as a constraint region
LASSO
RIDGE REGRESSION

The visualization above depicts what happens when we apply the two different regularization:

- The general idea is that we are restricting the allowed values of our coefficients to a certain "region" and within that region, we wish to find the coefficients that result in the best model

In this diagram,

- we are **fitting** a linear regression model with two features, x1 and x2
- β^ represents the set of two coefficients, β1 and β2, which minimize the RSS for the **unregularized model.**
- What does the ellipse represent?
  - The ellipses that are centred around β^ represent regions of constant RSS.
  - In other words, all of the points on a given ellipse share a common value of the RSS, despite the fact that they may have different values for β1 and β2 .
  - As the ellipses expand away from the least squares coefficient estimates, the RSS increases.
- What Regularisation does for the same model?
  - Regularization restricts the allowed positions of β^ to the blue constraint region.
  - In this case, β^ is not within the blue constraint region. Thus, we need to move β^ until it intersects the blue region, while increasing the RSS as little as possible.
- Region of circle/diamond

Ridge	Lasso

circle because it constrains the square of the coefficients. Thus the intersection will not generally occur on an axis, and so the coefficient estimates will be typically be **non-zero**.	- **diamond** because it constrains the absolute value of the coefficients. - Because the constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. - When this occurs, **one of the coefficients** will equal **zero**. - In higher dimensions, many of the coefficient estimates may equal zero simultaneously. In the figure above, the intersection occurs at $\beta1=0$, and so the resulting model will only include $\beta2$.

- size of the blue constraint region:
  - determined by λ , with a **smaller** λ resulting in a **larger region**
  - When λ is **zero**, the blue region is **infinitely large**, and thus the coefficient sizes are not constrained.
  - When λ increases, the blue region gets smaller and smaller.

Consideration before applying regularization

05 October 2022 00:09

Signs and causes of overfitting in the original regression model:

- Linear models can overfit if you include **irrelevant features**, meaning features that are unrelated to the response.
- Or if you include **highly correlated features**, Because it will learn a coefficient for every feature you include in the model, regardless of whether that feature has the signal or the noise. This is especially a problem when p (number of features) is close to n (number of observations).
- Linear models that have **large estimated coefficients** is a sign that the model may be overfitting the data.
- The larger the absolute value of the coefficient, the more power it has to change the predicted response, resulting in a higher variance.

Should features be standardized?

- **Yes**
- Because L1 and L2 regularizes of linear models assume that all features are centred around 0 and have variance in the same order.
- If a feature has a variance that is orders of magnitude larger than others, features would be penalized simply because of their scale and make the model unable to learn from other features correctly as expected.
- Also, standardizing avoids penalizing the intercept, which wouldn't make intuitive sense.

How should we choose between Lasso regression (L1) and Ridge regression (L2)?

- If model performance is our primary concern or that we are not concerned with explicit feature selection, it is best to try both and see which one works better. **Usually L2 regularization** can be expected to give superior performance over L1.
- Note that there's also a **ElasticNet regression**, which is a combination of Lasso regression and Ridge regression.
- Lasso regression is preferred if we want a sparse model, meaning that we believe many features are irrelevant to the output.
- When the dataset includes **collinear** features, **Lasso** regression is **unstable** in a similar way as unregularized linear models are, meaning that the coefficients (and thus feature ranks) can vary significantly even on small data changes
- When using L2-norm, since the coefficients are squared in the penalty expression, it has a different effect from L1-norm, namely it forces the coefficient values to be spread out more equally.

- When the dataset at hand contains correlated features, it means that these features should get similar coefficients

Example:

  - Using an example of a linear model $Y=X1+X2$, with strongly correlated feature of $X1$ and $X2$,
  - then for L1-norm, the penalty is the same whether the learned model is $Y=1*X1+1*X2$ or $Y=2*X1+0*X2$.
  - In both cases the penalty is $2*\alpha$. For L2-norm, however, the first model's penalty is $12+12=2\alpha$, while for the second model is penalized with $22+02=4\alpha$.

- The effect of this is that models are much more stable (coefficients do not fluctuate on small data changes as is the case with unregularized or L1 models).
- So while L2 regularization does not perform feature selection the same way as L1 does, it is more useful for feature interpretation due to its stability and the fact that useful features still tend to have non-zero coefficients.
- But again, please do remove collinear features to prevent a bunch of downstream headaches.