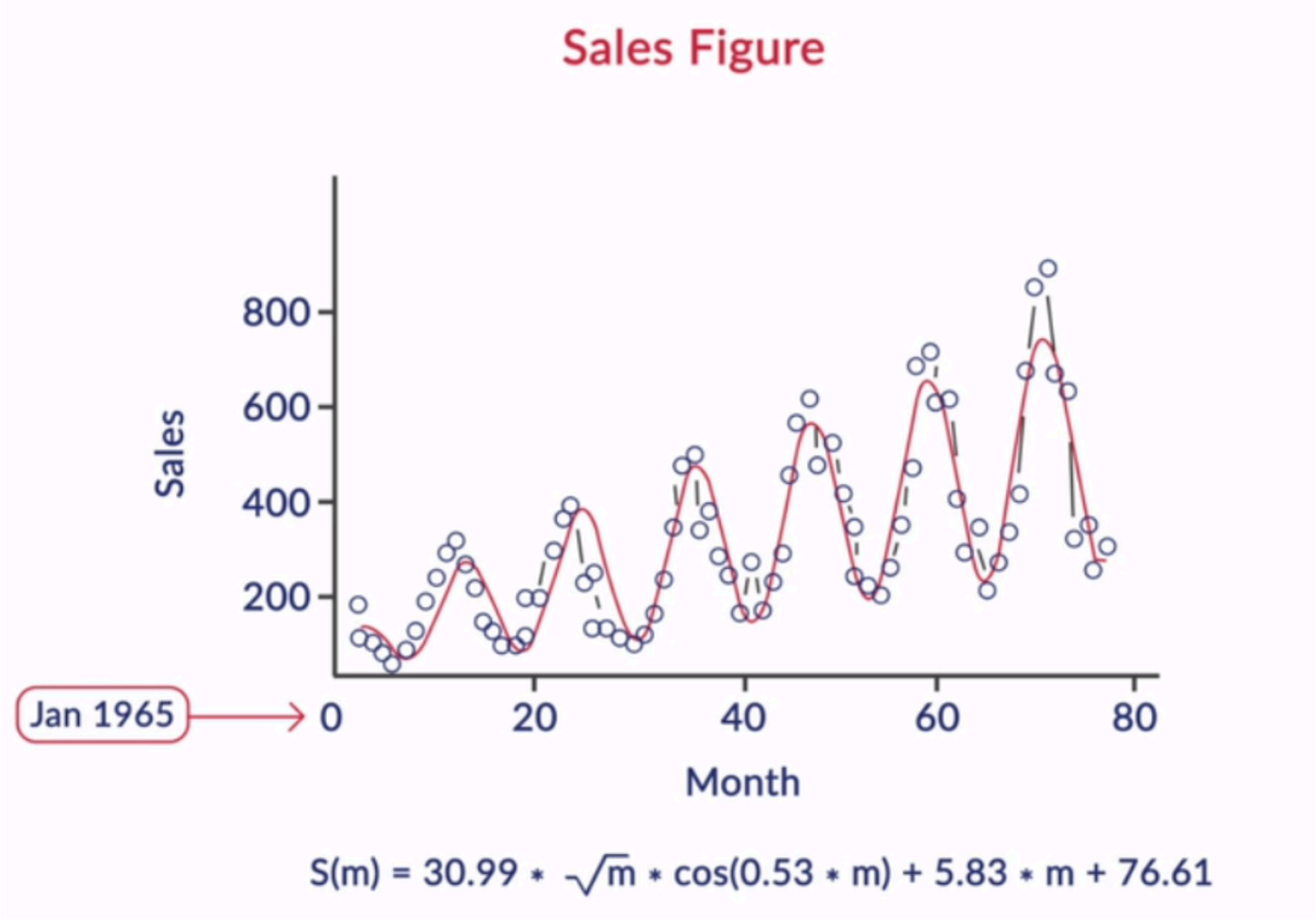
Lecture Notes

Advanced Regression

In this module, you were introduced to the concepts of the advanced regression framework. You have learnt to deal with the problems when the target variable y was non-linearly related to the predictor variables X. You were introduced to the concept of regularization in regression models. We discussed at length about the two regularized regression models, namely Ridge and Lasso. The concept of hyperparameter (λ) was also described in the context of regularization, along with its impact on the built model.

Generalized Regression

In linear regression, you had encountered problems where the target variable y was linearly related to the predictor variables X. But what if the relationship is not linear? Let's see how we can use **generalised regression** to tackle such problems.



© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

You should follow these two steps while building any model:

1. Carry out exploratory data analysis by examining scatter plots of explanatory and dependent variables.
2. Choose an appropriate set of functions which seem to fit the plot well, build models using them, and compare the results.

Feature Engineering

While constructing the non-linear regression model, instead of using the raw explanatory variables in the current form, we create some function of the explanatory variables to best explain the data points. These functions capture the non-linearity in the data

The derived features could be **combinations of two or more attributes** and/or **transformations of individual attributes**. These combinations and transformations could be **linear or non-linear.**

Note that a **linear combination** of two attributes x1 and x2 allows only two operations - **multiplying by a constant** and **adding the results**. For example, $3x_1 + 5x_2$ is a linear combination, though $2x_1x_2$ is a non-linear combination.

Generalized Regression Framework

We also saw several commonly used functions used in regression and how an n-degree polynomial can be expressed as a linear combination of features.

The next step is to find out the coefficients of such models mathematically, i.e. to fit the model. Let's see how we can do that.
In generalised regression models, the basic algorithm remains the same as linear regression- we compute the values of constants which result in the least possible error (best fit). The only difference is that we now use the features

$$\phi_1(x), \phi_2(x), \phi_3(x) \ldots \phi_k(x)$$

instead of the raw attributes.
The term 'linear' in linear regression refers to the linearity in the coefficients, i.e. the target variable y is **linearly related to the model coefficients**. It does not require that y should be linearly related to the raw attributes or features. Feature functions could be linear or non-linear.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

In a linear combination of features, the following operations can be performed:

1. We can multiply

$$\phi_1(x), \phi_2(x), \phi_3(x) \ldots . \phi_k(x)$$

with constants, for example,

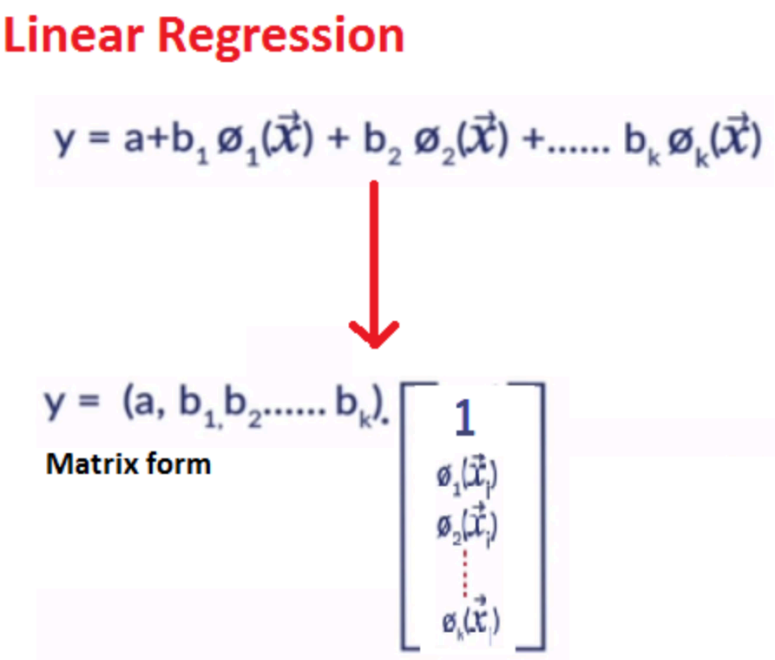
$$c_1\phi_1(x), c_2\phi_2(x), c_3\phi_3(x) \ldots . c_k\phi_k(x)$$

2. We can add those terms (but not multiply, divide, exponentiate etc.)
for example,

$$y = c_1\phi_1(x) + c_2\phi_2(x) + c_3\phi_3(x) \ldots . + c_k\phi_k(x)$$

Expressions

We can express the regression equation as a dot product of 2 vectors - 1. a vector of all the coefficients and 2. a vector with the features:



Next, we sum up the errors between predicted and actual response variables and minimize the residual sum of errors to get the optimal coefficients:

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

$$\underset{a,\, b_1,\, b_2 \ldots\ldots\, b_k}{\text{Min}} \sum_{i=1}^{n} \left(y_i - (a, b_1, b_2 \ldots\ldots\, b_k) \begin{bmatrix} 1 \\ \phi_1(\vec{x}_i) \\ \phi_2(\vec{x}_i) \\ \vdots \\ \phi_k(\vec{x}_i) \end{bmatrix} \right)^2$$

To summarise the key points:

1. We first created a feature matrix of dimension n x k, where n is the number of data points in the training dataset and k is the number of features.
2. We then use the expression to identify the coefficients that would correspond to the best-fit regression model and minimize the residual sum of errors:

$$(a, b_1, b_2 \ldots\ldots\, b_k) = (X^T X)^{-1} X^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

where,

$(X^T X)^{-1}$ = Inverse of matrix $(X^T X)$

X^T = Transpose of matrix X

$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ = Vector of observed values

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

As our goal is to minimise the loss function hence we'll differentiate the Loss function and equate it to zero.

$$L = \frac{1}{2}\sum_{t}^{N}(y^t - w^T x^t)^2 = \frac{1}{2}\left\|y - Xw\right\|^2 = \frac{1}{2}(y - Xw)^T(y - Xw)$$

$$Objective => \min_{w} L$$

$$\min_{w} \frac{1}{2}y^T y - w^T X^T y + \frac{1}{2}w^T X^T Xw$$

$$\frac{dL}{dw} = 0 = -X^T y + X^T Xw$$

$$w = (X^T X)^{-1} X^T y$$

Regularized Regression

A predictive model has to be as simple as possible, but no simpler. There is an important relationship between the complexity of a model and its usefulness in a learning context because of the following reasons:

- Simpler models are usually more generic and are more widely applicable (are generalizable)
- Simpler models require fewer training samples for effective training than the more complex ones

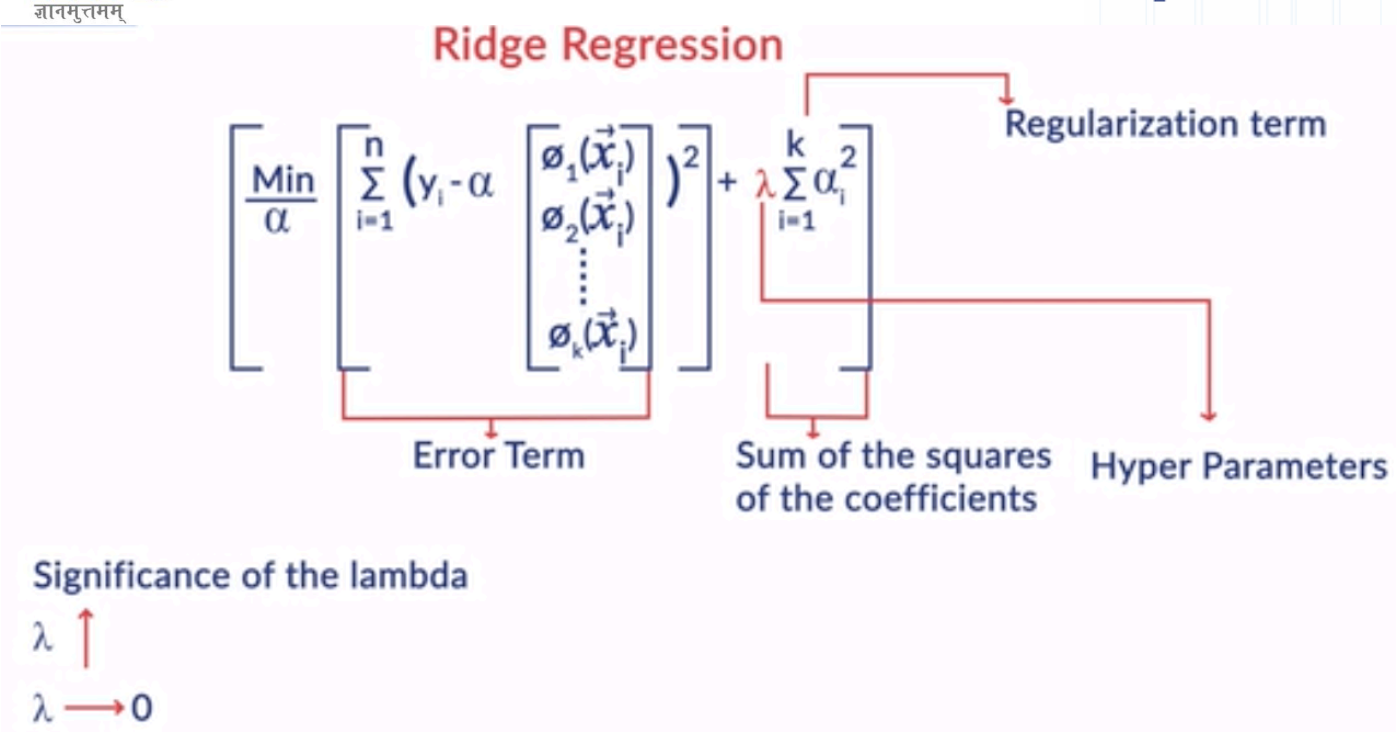
Regularization is a process used to create an optimally complex model, i.e. a model which is as simple as possible while performing well on the training data.

Through regularization, the algorithm designer tries to strike the delicate balance between keeping the model simple, yet not making it too naive to be of any use.
The regression does not account for model complexity - it only tries to minimize the error (e.g. MSE), although if it may result in arbitrarily complex coefficients. On the other hand, in regularized regression, the objective function has two parts - the **error term** and the **regularization term**.

Ridge Regression

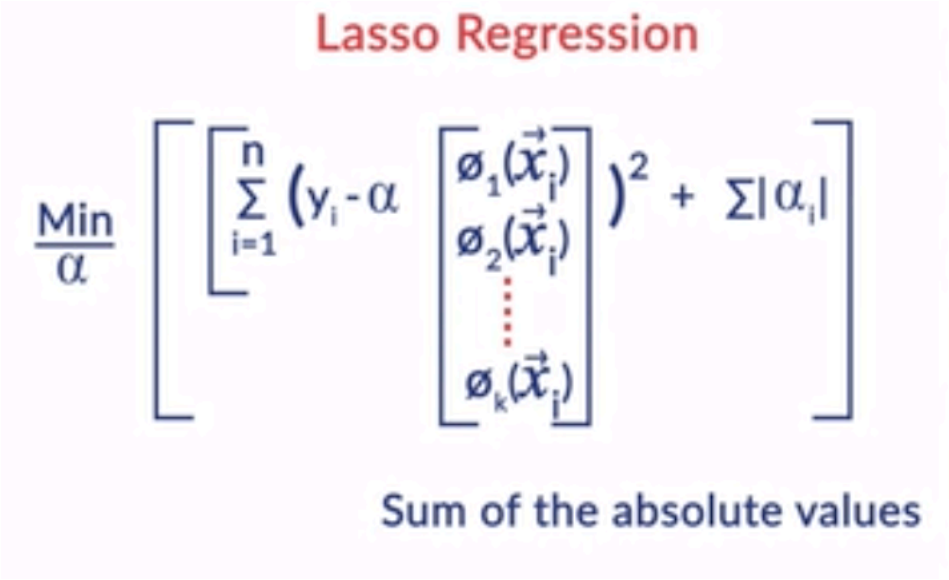
In ridge regression, an additional term of "sum of the squares of the coefficients" is added to the cost function along with the error term

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved



Lasso Regression

In case of lasso regression, a regularisation term of "sum of the absolute value of the coefficients" is added



These are two commonly used regularised regression methods - Ridge regression and Lasso regression. Both these methods are used to make the regression model simpler while balancing the 'bias-variance' trade-off.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

Difference between Ridge and Lasso Regression

You learnt that both Ridge and Lasso regularize the coefficients by reducing them in value, essentially causing shrinkage of the coefficients. Ridge and Lasso perform different measures of shrinkage which depends on the value of hyperparameter, λ. In the process of shrinkage, **Lasso shrinks some of the variable coefficients to 0**, thus performing variable selection.

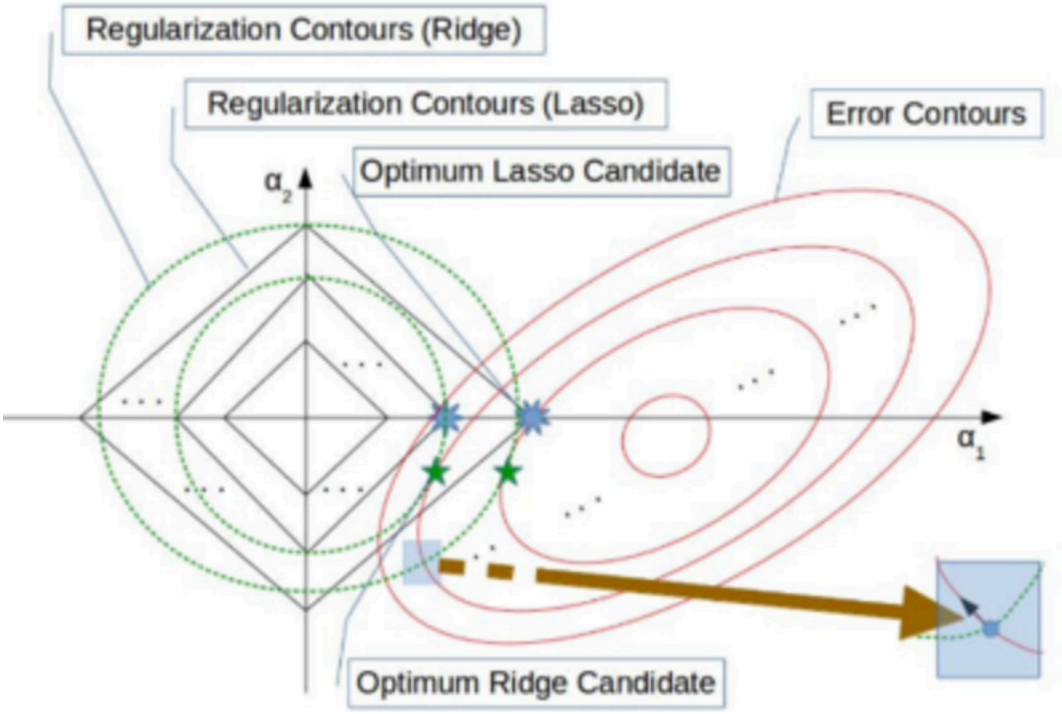


Figure 9: Lasso vs Ridge Regression

Thus, the key observation here is that at the optimum solution for α (the place where the sum of the error and regularisation terms is minimum), the corresponding regularization contour and the error contour must 'touch' each other tangentially and not 'cross'. The 'blue stars' highlight the touch points between the error contours and the lasso regularization contours. The 'green stars' highlight the touch points between the error contours and the ridge regularization terms. The picture illustrates the fact that because of the 'corners' in the lasso contours (unlike ridge regression), the touch points are more likely to be on one or more of the axes. This implies that the other coefficients become zero. Hence, lasso regression also serves as a variable shrinkage method, whereas ridge regression does not.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

Model Selection Parameters

While creating the best model for any problem statement, we end up choosing from a set of models which would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order to select the best model. This can be done in the following two ways.

1. Use metrics which take into account both model fit and simplicity. They penalise the model for being too complex (i.e. for overfitting), and thus are more representative of the unseen 'test error'. Some examples of such metrics are Mallow's Cp, Adjusted R^2, AIC and BIC
2. Estimate the test error via a validation set or a cross-validation approach.
   In validation set approach, we find the test error by training the model on training set and fitting on an unseen validation set while in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold where k varies from 1 to n.

Let's look into these one by one

1. Mallow's Cp

$$C_p = \frac{1}{n}(RSS + 2d\sigma^2)$$

2. AIC (Akaike information criterion)

$$AIC = \frac{1}{n\sigma^2}(RSS + 2d\sigma^2)$$

3. BIC (Bayesian information criterion)

$$BIC = \frac{1}{n}(RSS + ln(n)d\sigma^2)$$

4. Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

AIC and BIC are defined for models fit by maximum likelihood estimator. We can notice that as we increase the number of predictors d, the penalty term in Cp, AIC and BIC all increase while the RSS decreases. Hence, lower the value of Cp, AIC and BIC, better is the fit of the model. Higher the Adjusted R^2 , better is the fit of the model.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

Best Subset Selection

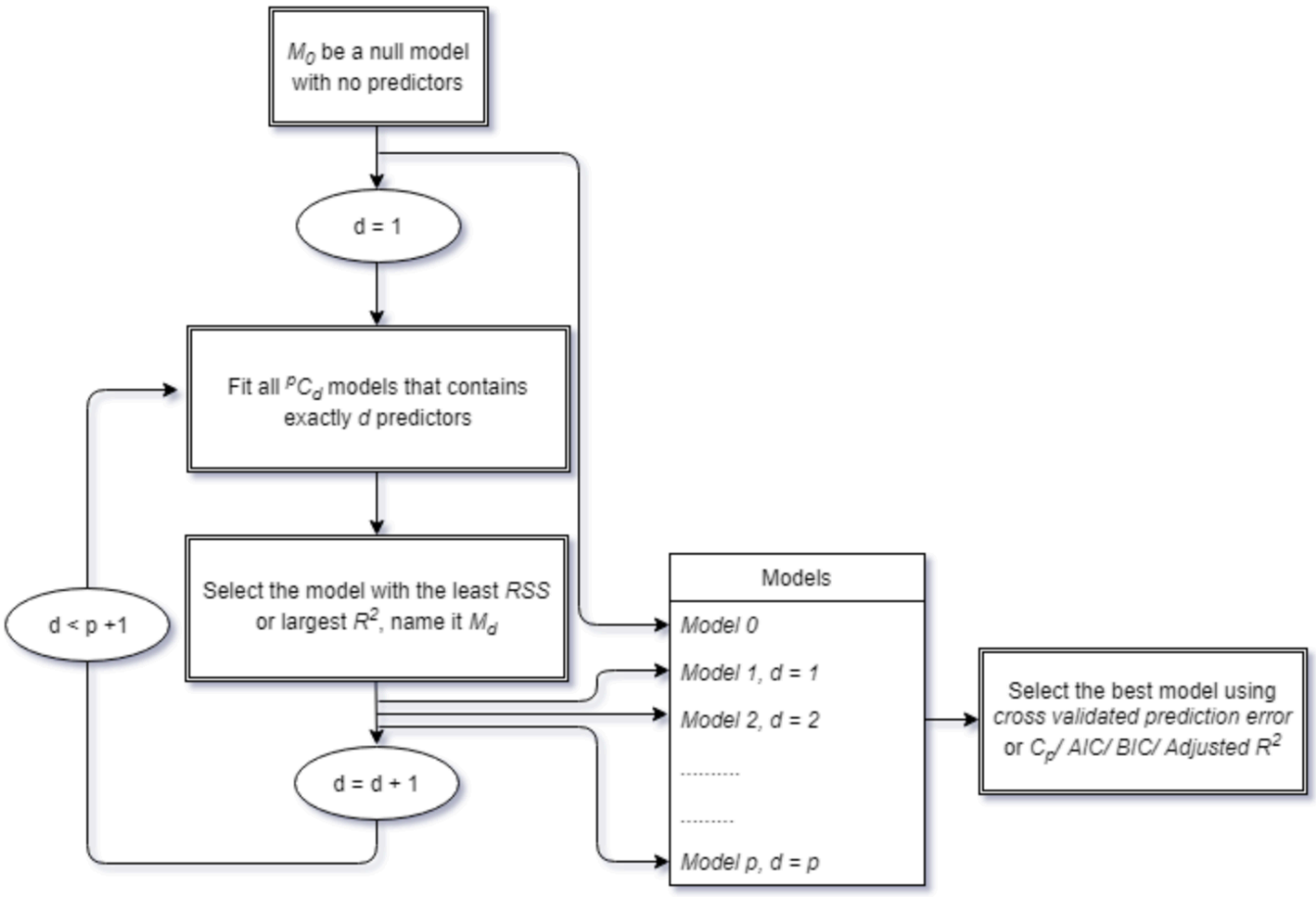
Now, we will look at the different methods of choosing the best set of predictors that shall give the least test error.

Features' subset selection can be performed using two different methods:

1. Best Subset Selection

A brief explanation of the Best Subset Selection algorithm (run on a dataset with p features) is as follows (please refer to the image below): You start with d=0 features, i.e. a null model M0 with no features. Now, as you increase d, you consider every model that has all combinations of d features and select a model which results in the least RSS (or largest R2). This gives you a model Md with d features. Continue this iteration by increasing the value of d by one till you reach d=p and find the models M0, M1, M2,.....,Mp.

Out of all these models M0, M1, M2,.....,Mp, select the best one, as measured by a measure such as Cp, AIC, BIC, Adjusted R^2 or mean cross-validated error.



We can see that the total number of models that need to be analysed for Best Subset Selection is 2^p where p is the total number of predictors. If we have 20 predictors, the total number of models is 2^{20} = 1048576 which is over a million. Hence, it becomes computationally infeasible to perform best subset selection for number of predictors greater than 40.

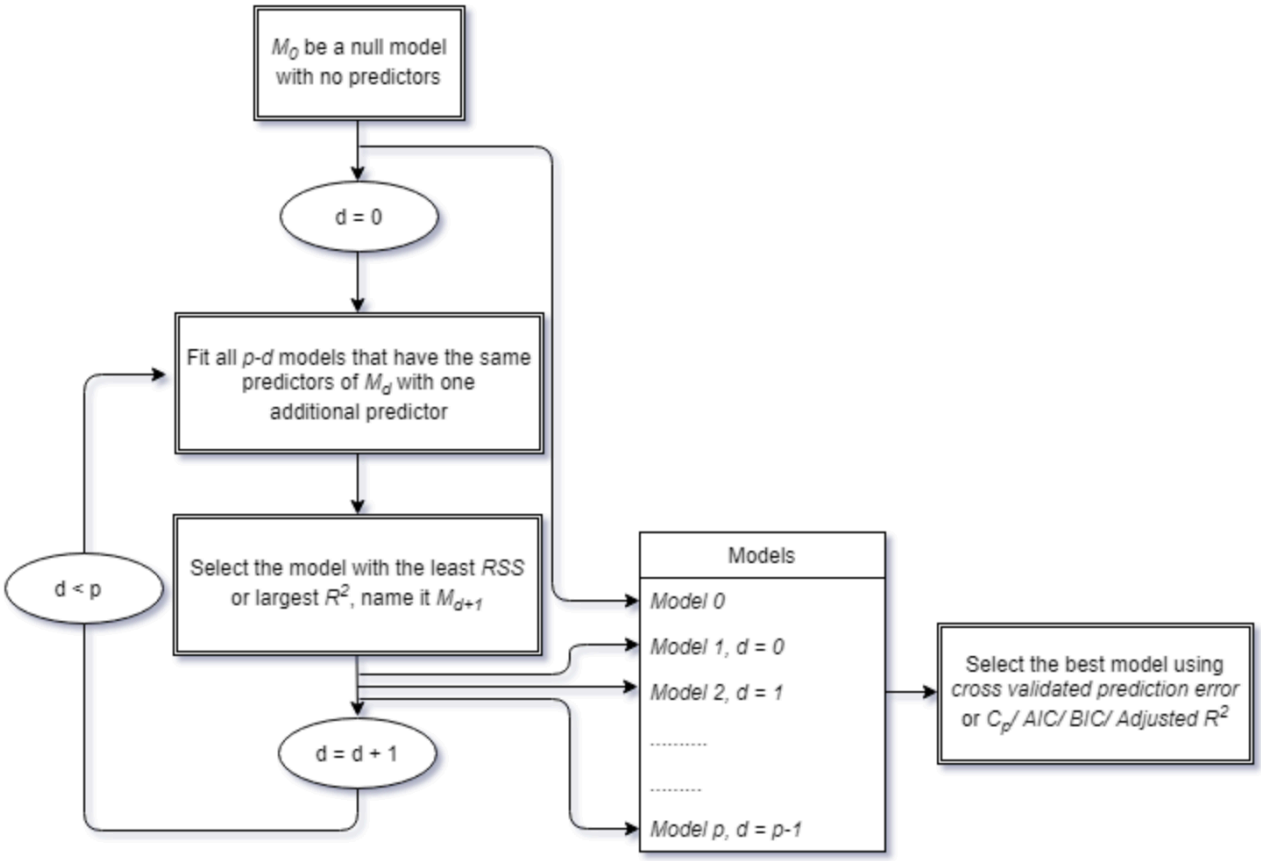
© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

2. Stepwise Selection

Forward Stepwise Selection

A brief explanation of the forward selection algorithm (run on a dataset with p features) is as follows (please refer to the flowchart below): You start with d=0 features, i.e. a null model M0 with no features. Now, out of the (p-d) remaining features, you identify one additional feature which (when added to the model Md) results in the least RSS (or largest R2). This gives you a model Md+1 with one additional feature. Continue this iteration by increasing the value of d by one till you reach d=p and find the models M0, M1, M2,.....,Mp.

Out of all these models M0, M1, M2,.....,Mp, select the best one, as measured by a measure such as Cp, AIC, BIC, Adjusted R^2 or mean cross-validated error.

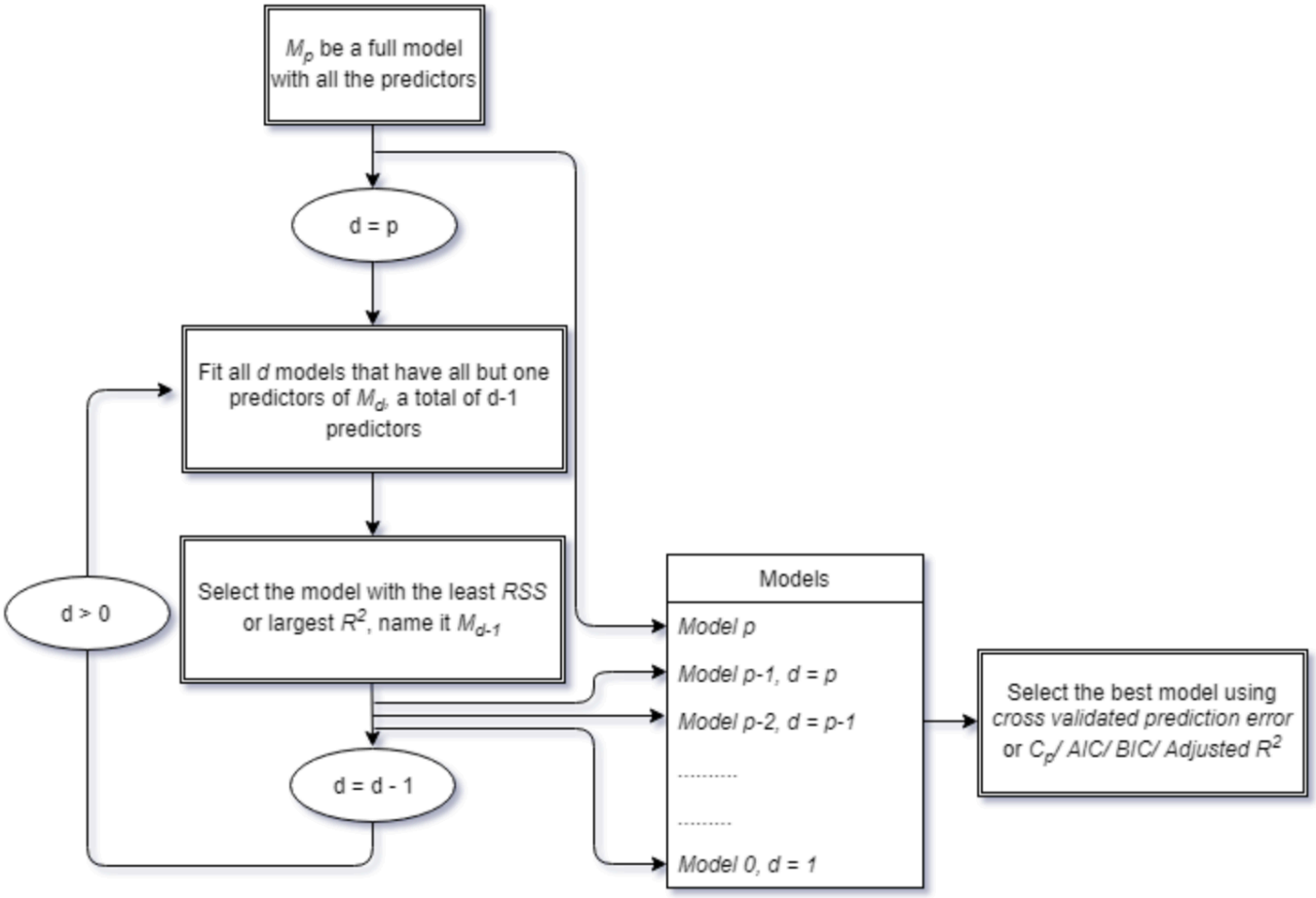


Backward Stepwise Selection

The backward selection algorithm is the opposite of the forward one - rather than starting with d=0 features and adding a feature in each iteration, you start with d=p features (a model Mp, with all the features as predictors) and remove a feature in every iteration - the one that minimises the error (or maximises R2) and find the models Mp, Mp−1, Mp−2,....., M0.

Out of all these models Mp, Mp−1, Mp−2,....., M0, select the best one, as measured by a measure such as Cp, AIC, BIC, Adjusted R^2 or mean cross-validated error.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved



We can see that the total number of models that need to be analysed for Forward Stepwise Selection is 1+p(p+1)/2 where p is the total number of predictors. It is the same for Backward Stepwise Selection also.

So, if the number of predictors is 40, there are just 821 models that need to analysed which is significantly lesser than 2^{40}. In this way, it is better than Best Subset Selection but it also has limitations.

Stepwise Selection does not ensure that we have chosen the best model. If we start off Forward Stepwise Selection with predictor X1 then the best model with 2 predictors becomes X1 and X2 since X2 is the next best predictor. But the best model with 2 predictors may be X2 and X3. This can happen with Backward Stepwise Selection also.

Though, Forward Stepwise Selection can be applied for n<p, where n is the number of observations, Backward Stepwise Selection cannot be applied, as a full model cannot be fit when n<p.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved

Disclaimer: All content and material on the UpGrad website is copyrighted material, either belonging to UpGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of UpGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document will be modified, adapted or altered in any way.
- No part of this document or UpGrad content may be reproduced or stored in any other web site or included in any public or private electronic retrieval system or service without UpGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights reserved