

# MACHINE LEARNING

**In Q1 to Q11, only one option is correct, choose the correct option:**

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?  
A) Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

**Ans A**

2. Which of the following statement is true about outliers in linear regression?

A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers C) Can't say  
D) none of these

**Ans A**

3. A line falls from left to right if a slope is \_\_\_\_\_?

A) Positive B) Negative C) Zero D) Undefined

**Ans B**

4. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression B) Correlation C) Both of them D) None of these

**Ans A**

5. Which of the following is the reason for over fitting condition?

A) High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these

**Ans C**

6. If output involves label then that model is called as:

A) Descriptive model B) Predictive modal C) Reinforcement learning D) All of the above

**Ans A**

7. Lasso and Ridge regression techniques belong to \_\_\_\_\_?

A) Cross validation B) Removing outliers C) SMOTE D) Regularization

**Ans D**

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation B) Regularization C) Kernel D) SMOTE

**Ans D**

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?

A) TPR and FPR B) Sensitivity and precision C) Sensitivity and Specificity D) Recall and precision

**Ans A**

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True B) False

**Ans A**

11. Pick the feature extraction from below:

A) Construction bag of words from a email B) Apply PCA to project high dimensional data C) Removing stop words D) Forward selection

**Ans A**

**In Q12, more than one options are correct, choose all the correct options:**

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

A) We don't have to choose the learning rate. B) It becomes slow when number of features is very large. C) We need to iterate. D) It does not make use of dependent variable.

**Ans B, C**

**Q13 and Q15 are subjective answer type questions, Answer them briefly.**

13. Explain the term regularization?


**Ans** Regularization

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, ***this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.***

A simple relation for linear regression looks like this. Here Y represents the learned relation and  $\beta$  represents the coefficient estimates for different variables or predictors(X).

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

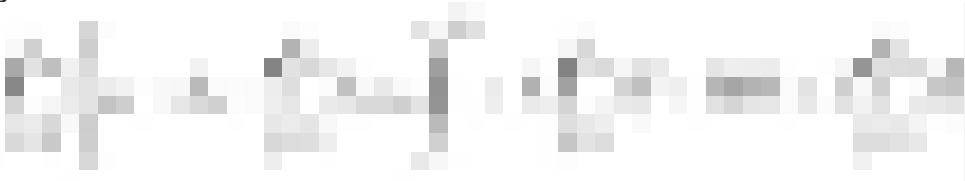
The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.



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

Now, this will adjust the coefficients based on your training data. *If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.*

#### Ridge Regression



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

Above image shows ridge regression, where the ***RSS is modified by adding the shrinkage quantity.*** Now, the coefficients are estimated by minimizing this function. Here,  ***$\lambda$  is the tuning parameter that decides how much we want to penalize the flexibility of our model.*** The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept  $\beta_0$ , This intercept is a measure of the mean value of the response when  $x_{i1} = x_{i2} = \dots = x_{ip} = 0$ .

When  $\lambda = 0$ , the penalty term has no effect, and the estimates produced by ridge regression will be equal to least squares. However, **as  $\lambda \rightarrow \infty$ , the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.** As can be seen, selecting a good value of  $\lambda$  is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are **also known as the L2 norm.**

**The coefficients that are produced by the standard least squares method are scale equivariant**, i.e. if we multiply each input by  $c$  then the corresponding coefficients are scaled by a factor of  $1/c$ . Therefore, regardless of how the predictor is scaled, the multiplication of predictor and coefficient ( $X_j\beta_j$ ) remains the same. **However, this is not the case with ridge regression, and therefore, we need to standardize the predictors or bring the predictors to the same scale before performing ridge regression.** The formula used to do this is given below.

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

Lasso

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

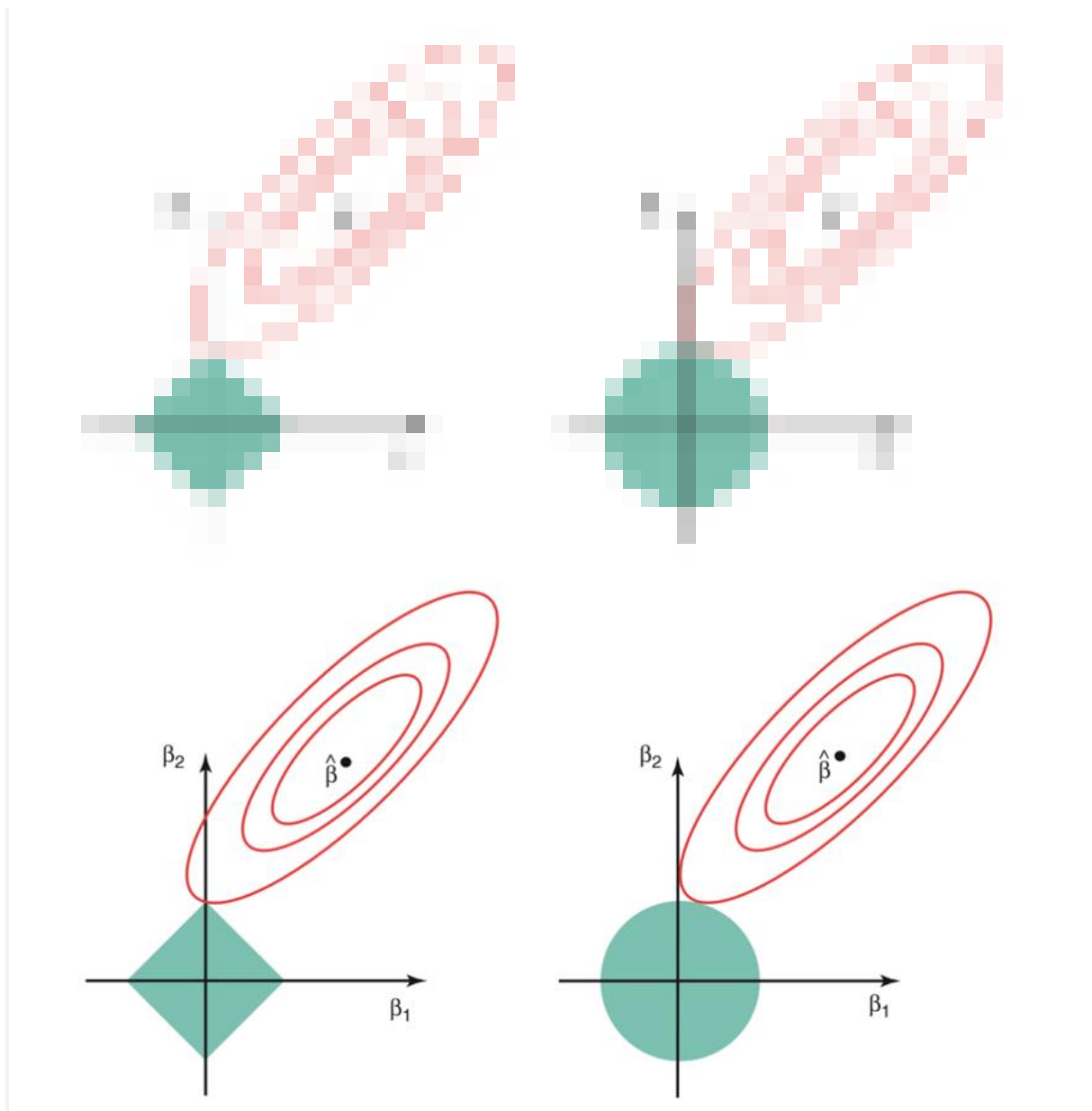
Lasso is another variation, in which the above function is minimized. Its clear that ***this variation differs from ridge regression only in penalizing the high coefficients***. It uses  $|\beta_j|$  (modulus) instead of squares of  $\beta$ , as its penalty. In statistics, this is ***known as the L1 norm***.

Lets take a look at above methods with a different perspective. *The ridge regression can be thought of as solving an equation, where summation of squares of coefficients is less than or equal to s. And the Lasso can be thought of as an equation where summation of modulus of coefficients is less than or equal to s.* Here, s is a constant that exists for each value of shrinkage factor  $\lambda$ . ***These equations are also referred to as constraint functions.***

***Consider their are 2 parameters in a given problem.*** Then according to above formulation, the ***ridge regression is expressed by  $\beta_1^2 + \beta_2^2 \leq s$*** . This implies that *ridge regression coefficients have the smallest RSS(loss function) for all points that lie within the circle given by  $\beta_1^2 + \beta_2^2 \leq s$ .*

Similarly, ***for lasso, the equation becomes,  $|\beta_1| + |\beta_2| \leq s$*** . This implies that *lasso coefficients have the smallest RSS(loss function) for all points that lie within the diamond given by  $|\beta_1| + |\beta_2| \leq s$ .*

The image below describes these equations.



Credit: An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

***The above image shows the constraint functions (green areas), for lasso(left) and ridge regression(right), along with contours for RSS (red ellipse).*** Points on the ellipse share the value of RSS. For a very large value of  $s$ , the green regions will contain the centre of the ellipse, making coefficient estimates of both regression techniques, equal to the least squares estimates. But, this is not the case in the above image. In this case, the lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint

region. *Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. However, the lasso 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 (where parameters are much more than 2), many of the coefficient estimates may equal zero simultaneously.

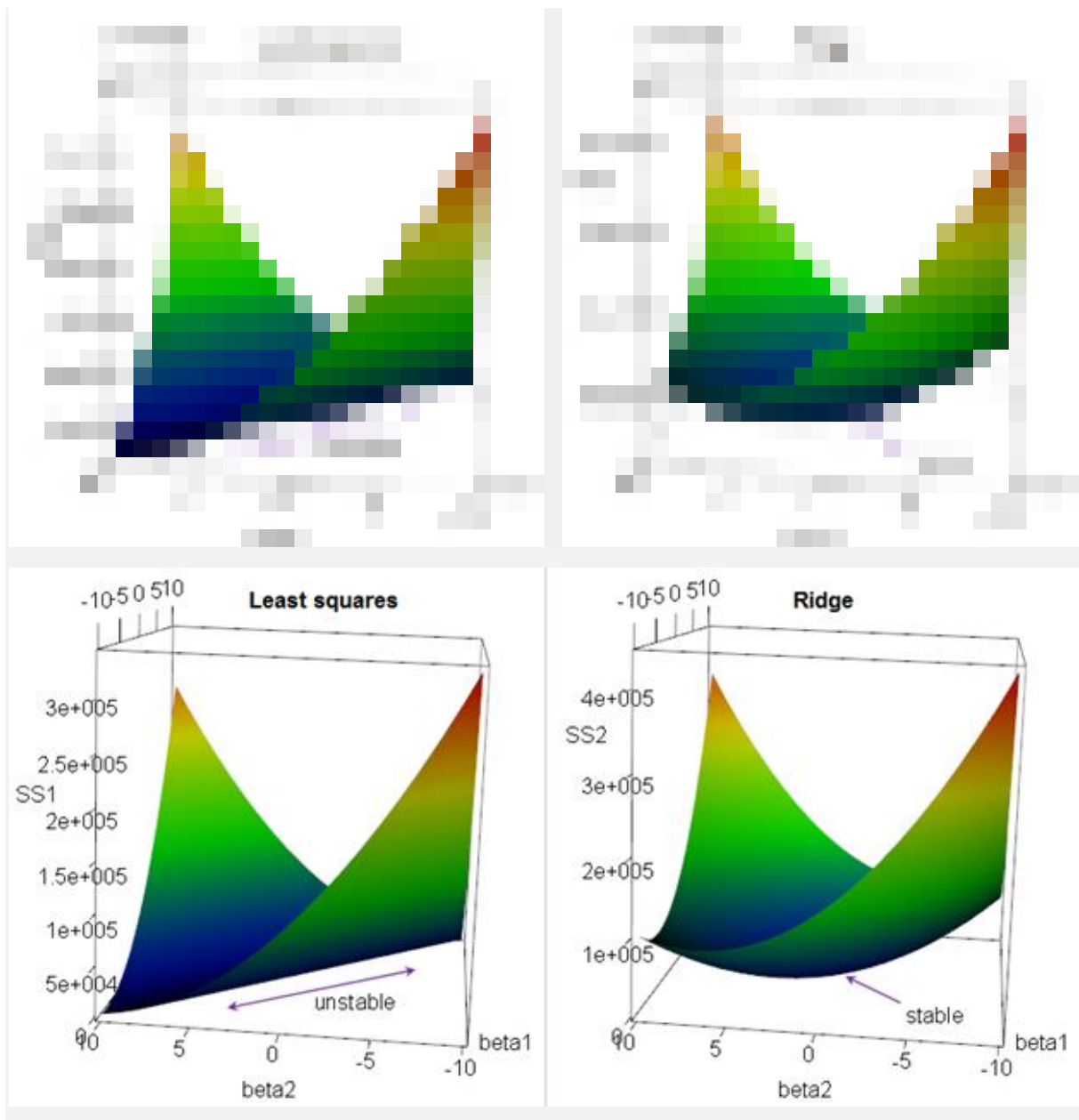
14. Which particular algorithms are used for regularization?

Ans **Regularization Algorithms:**

I find these algorithms particular useful. Most data scientists will find some of their models to overfit at some point during their career. The general idea behind these algorithms is that they try to minimize and even prevent overfitting.

- **Ridge Regression (L2 Regularization)**

Its goal is to solve problems of data overfitting and when the data suffers from multicollinearity (Multicollinearity in a multiple regression model are highly linearly related associations between two or more explanatory variables). A standard linear or polynomial regression model will fail in the case where there is high collinearity (the existence of near-linear relationships among the independent variables) among the feature variables. Ridge Regression adds a small squared bias factor to the variables. Such a squared bias factor pulls the feature variable coefficients away from this rigidness, introducing a small amount of bias into the model but greatly reducing the variance.



*Some things to consider:*

Ridge works very well to avoid over-fitting.

If you have a model with a large number of features in the dataset and you want to avoid making the model too complex, use regularization to address over-fitting and feature selection.



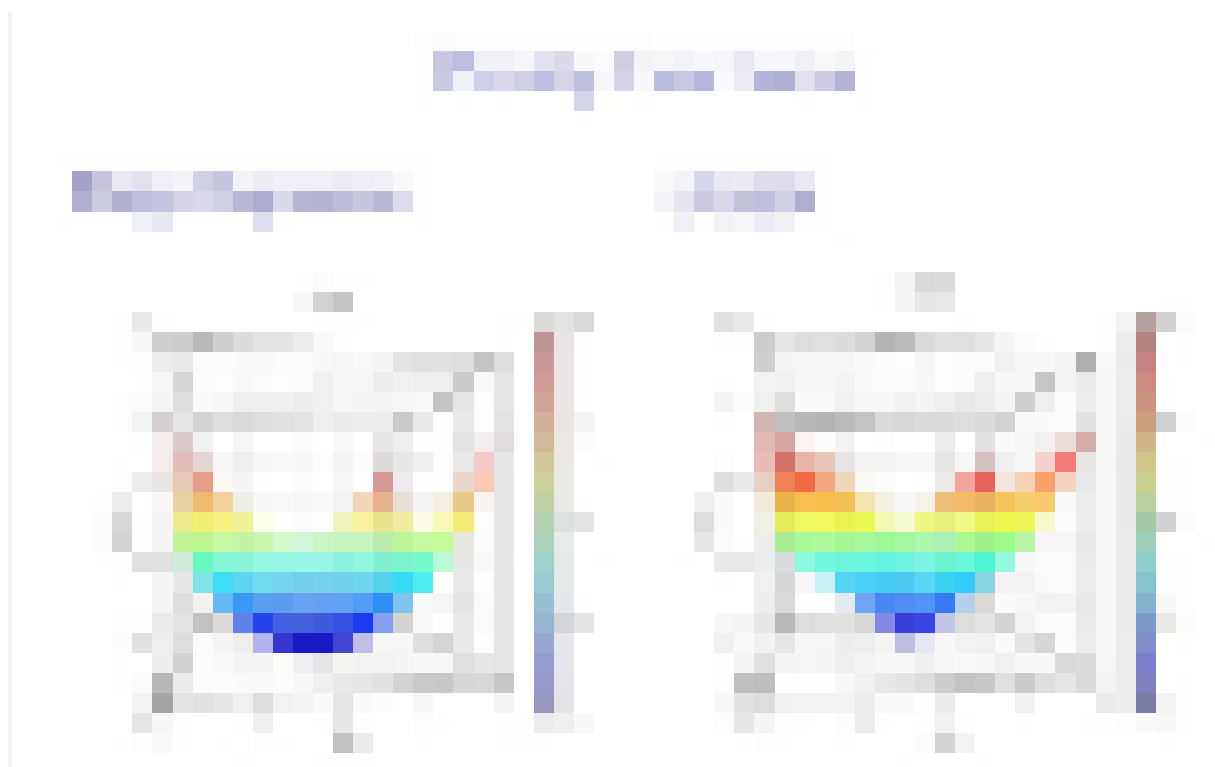
Ridge has one main disadvantage, it includes all  $N$  features in the final model.

When you have highly-correlated variables, Ridge regression shrinks the two coefficients towards one another. Lasso is somewhat indifferent and generally picks one over the other.

Depending on the context, one does not know which variable gets picked. Elastic-net is a compromise between the two that attempts to shrink and do a sparse selection simultaneously.

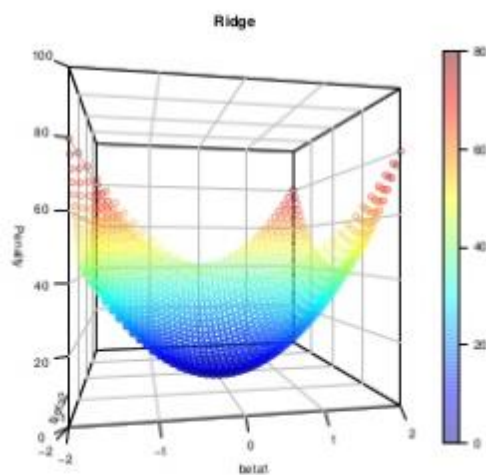
- **Least Absolute Shrinkage and Selection Operator (LASSO, L1 Regularization)**

In opposite to Ridge Regression it only penalizes high coefficients. Lasso has the effect of forcing some coefficient estimates to be exactly zero when hyper parameter  $\theta$  is sufficiently large. Therefore, one can say that Lasso performs variable selection producing models much easier to interpret than those produced by Ridge Regression. Basically, it is reducing the variability and improving the accuracy of linear regression models.

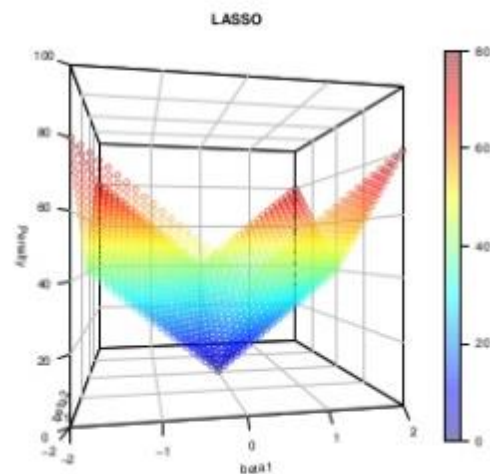


## Penalty Functions

### Ridge Regression



### LASSO



*Some things to consider:*

Lasso is a regularization technique for performing linear regression.

Lasso is one alternative method to stepwise regression and other model selection and dimensionality reduction techniques.

LASSO works well for feature selection in case we have a huge number of features (it reduce redundant features and identify the important ones).

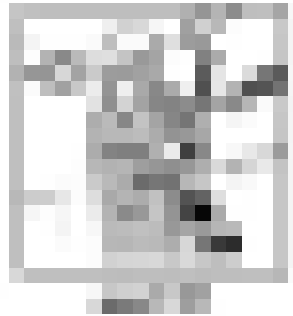
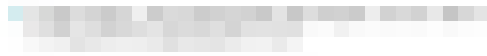
It shrinks coefficients to zero (compare to Ridge which adds “squared magnitude” of coefficient as penalty term to the loss function).

If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero.

Other methods like cross-validation, stepwise regression work fairly well for reducing overfitting and perform feature selection. However, they mainly work with a small amount of features. Ridge and LASSO work well with a large amount of features.

- Elastic Net

Combines characteristics of both lasso and ridge. Elastic Net reduces the impact of different features while not eliminating all of the features. Lasso will eliminate many features, and reduce overfitting in your linear model. Ridge will reduce the impact of features that are not important in predicting your y values. Elastic Net combines feature elimination from Lasso and feature coefficient reduction from the Ridge model to improve your model's predictions.



## Elastic Net

- ◆ Ridge, Lasso, and Elastic Net are all part of the same family with the penalty term of:

$$P_{\alpha} = \sum_{i=1}^p \left[ \frac{1}{2} (1 - \alpha) b_i^2 + \alpha |b_i| \right]$$

- ◆ If the  $\alpha = 0$  then we have a **Ridge Regression**
- ◆ If the  $\alpha = 1$  then we have the **LASSO**
- ◆ If the  $0 < \alpha < 1$  then we have the **elastic net**



*Some things to consider:*

Use elastic net when you have several highly correlated variables.

Useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

Studies have suggested that the elastic net technique can outperform LASSO when used on similar data with highly correlated predictors.

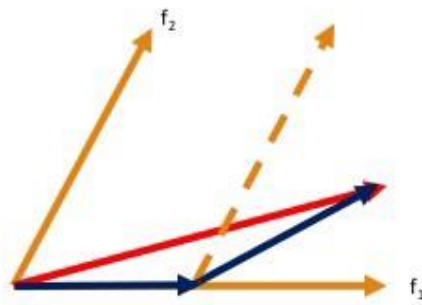
- Least-Angle Regression (LARS)

Similar to forward stepwise regression. At each step, it finds the predictor most correlated with the response. When multiple predictors having equal correlation exist, instead of continuing along the same predictor, it proceeds in a direction equiangular between the predictors. Least angle regression is like a more “democratic” version of forward stepwise regression. It follows the same general scheme of forward stepwise regression, but doesn’t add a predictor fully into the model. The coefficient of that predictor is increased only until that predictor is no longer the one most correlated with the residual  $r$ . Then some other competing predictor is invited to “join the club”. It start with all coefficients equal to zero, and then it finds the predictor that is most correlated with  $y$ . It increases the coefficient in the direction of the sign of its correlation with  $y$ , and then it’s taking residuals along the way and stopping when some other predictor has as much correlation with  $r$  as the first one has.



## LARS (Least Angle regression)

Move along most correlated feature until another feature becomes equally correlated



15. Explain the term error present in linear regression equation?

**Ans** An error term represents the margin of error within a statistical model; it refers to the **sum of the deviations within the regression** line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results.

The distance between each point and the linear graph (shown as black arrows on the above graph) is our error term. So we can write our function as  $R^B = \beta_0 + \beta_1 E^x + \varepsilon$  where  $\beta_0$  and  $\beta_1$  are constants and  $\varepsilon$  is an (non constant) error term.

