

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

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

**Answer: A)** Least Square Error

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

**Answer: A)** Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is ?
- A) Positive
  - B) Negative
  - C) Zero
  - D) Undefined

**Answer: B)** Negative

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

**Answer: C)** Both of them

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

**Answer: C)** Low bias and high variance

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

**Answer: B)** Predictive modal

7. Lasso and Ridge regression techniques belong to ?
- A) Cross validation
  - B) Removing outliers
  - C) SMOTE
  - D) Regularization

**Answer: D)** Regularization

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

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

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

**Answer: A) Cross Validation**

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

**Answer: A) TPR and FPR**

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

- A) True
- B) False

**Answer: B) False**

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

**Answer: C) Removing stop words**

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

1. 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.

**Answer: A, B, C**

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

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

**13. Explain the term regularization?**

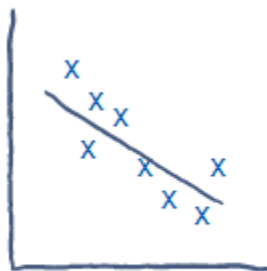
**Answer:**

14.

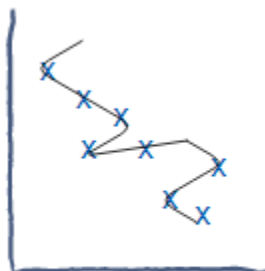
Regularization means to make things regular or acceptable. This is exactly why we use it for applied machine learning. In the context of machine learning, regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent over fitting.

**Definition of Regularization**

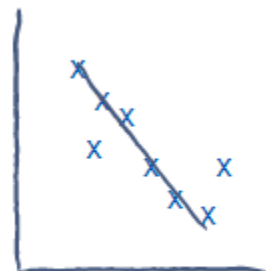
**Regularization** refers to training our model well enough that it can generalize over data it hasn't seen before. Regularization is a common method used to reduce over fitting and improve the model's performance for new inputs.



Underfitted model



Overfitted model



Fitted model

Regularization Works:

**Regularization methods** are techniques that seek to reduce overfitting (i.e., reduce generalization errors) by keeping network weights small. There are three very popular and efficient regularization techniques called L1, L2, and dropout.

**L1 regularization**

In the case of L1 regularization (also known as Lasso regression), we simply use another regularization term,  $\Omega$ . This term is the sum of the absolute values of the weight parameters in a weight matrix. L1 encourages weights to 0.0 (if possible), which results in more sparse weights (more weights with values equal to 0.0). Hence, the cost function in L1 becomes:

# MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} * \sum \|w\|$$

## L2 regularization

**L2 regularization** offers more nuances by penalizing larger weights more severely, thus resulting in weights that are less sparse. The regularization term ( $\Omega$ ) is defined as the Euclidean Norm (or L2 norm) of the weight matrices and is the sum over all squared weight values of a weight matrix. The cost function in L2 becomes:

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} * \sum \|w\|^2$$

## Dropout regularization

Dropout regularization involves a neuron of the neural network getting turned off during training with a probability of PP. This results in a simpler neural network since some neurons are not active at all.

A simpler version of the neural network results in less complexity, which can reduce overfitting. The deactivation of neurons with a certain probability (PP) is applied at each forward propagation and weight update step.

## 14. Which particular algorithms are used for regularization?

### Answer:

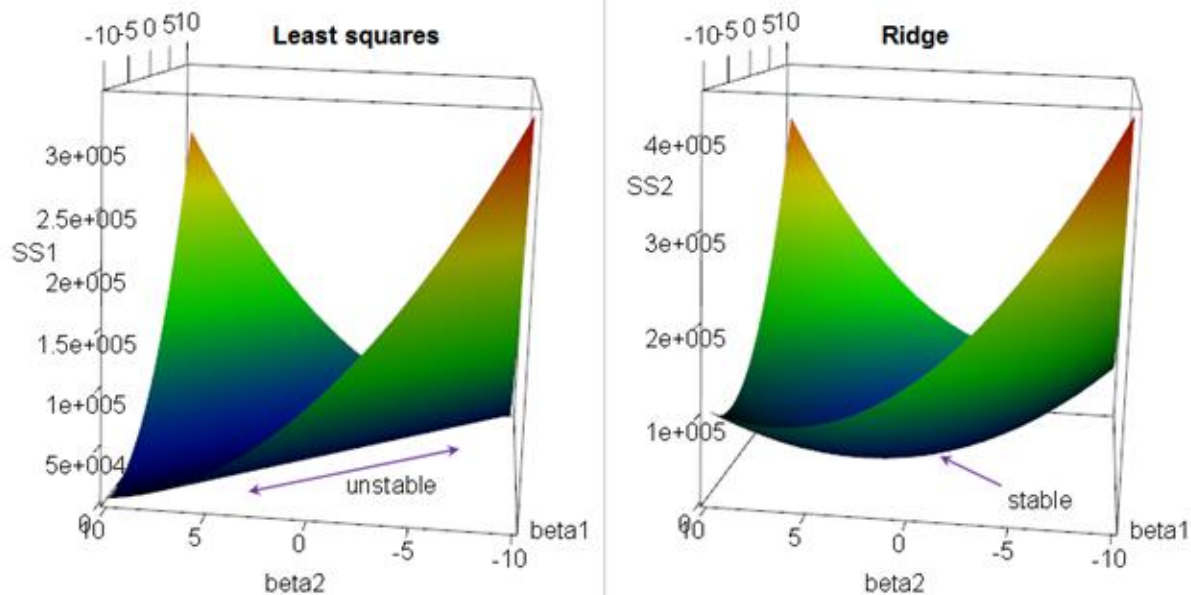
I find below mentioned algorithms particular useful for regularization. The general idea behind these algorithms is that to try minimizing and even preventing overfitting.

### 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

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

coefficients away from this rigidness, introducing a small amount of bias into the model but greatly reducing the variance.



Advantages:

1. Ridge works very well to avoid over-fitting.
2. 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.

Disadvantage:

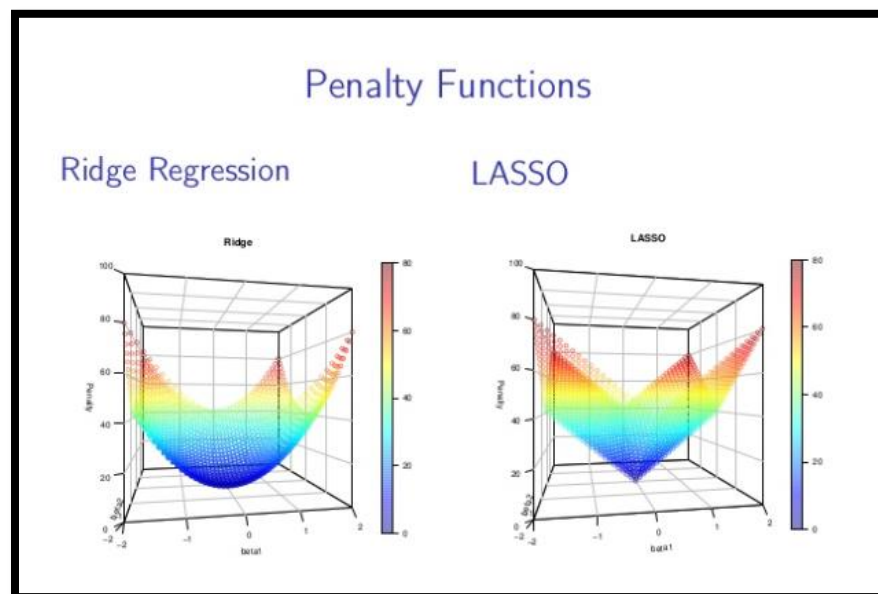
1. It includes all  $N$  features in the final model.

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

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

## 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.



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.

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

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



**PENALTY BOX**

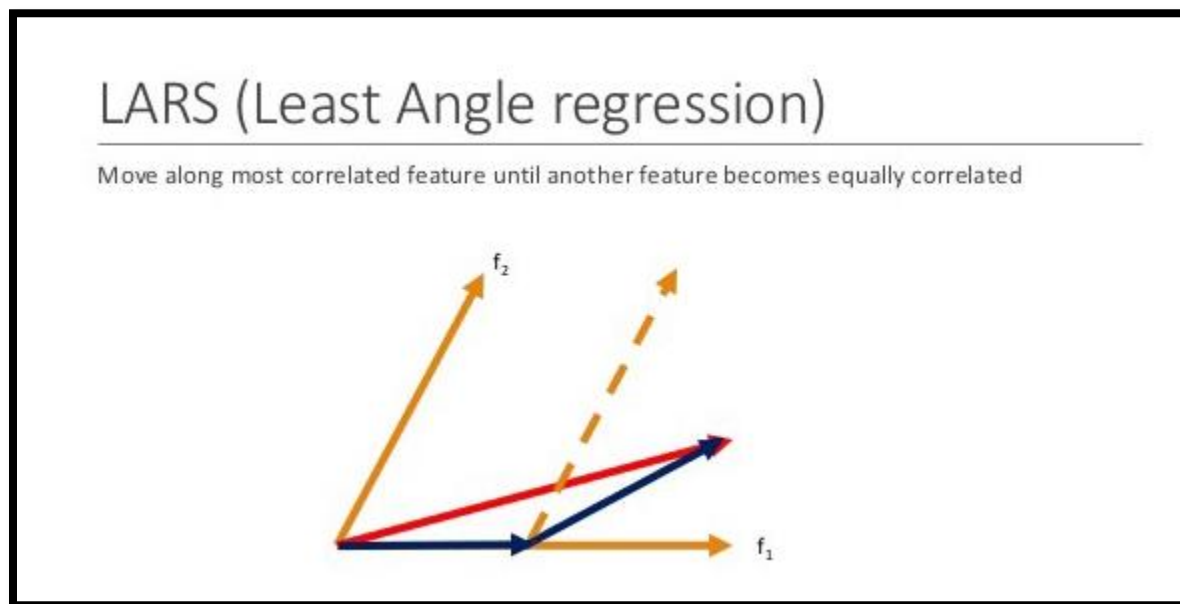
- Use elastic net when you have several highly correlated variables.
- Useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.
- Studies have suggested that the elastic net technique can outperform LASSO when 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

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

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.



- It is useful when the number of dimensions is significantly greater than the number of points
- If two variables are almost equally correlated with the response, then their coefficients should increase at approximately the same rate.

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

**Answer:**



## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

An error term is a variable in a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. ... The error term is also known as the residual, disturbance or remainder term.

Perfect prediction on the basis of regression equation seems to be difficult. What is needed is a measure that would indicate how precise the prediction of  $y$  is based on  $x$ . The term error in a regression model is a standard error of estimate. This standard error of estimate has the same concept as the standard deviation.

The standard deviation measures dispersion of the observations about the mean of distribution while the standard error of estimate measures dispersion about the line of regression.

In other words standard error measures the accuracy of estimated value. The smaller the value of standard error, lesser the dispersion around the regression. If the error value is zero it means there is no deviation about the line. That shows that the correlation is perfect.

### ASSUMPTION ABOUT ERROR TERM

1. The error  $\varepsilon$  is a random variable with mean of zero.
2. The variance of  $\varepsilon$ , denoted by  $\sigma^2$ , is the same for all values of the independent variable.
3. The values of  $\varepsilon$  are independent.
4. The error  $\varepsilon$  is a normally distributed random variable.

In the linear regression model, the statistical term "Standard Error of Regression" has to do with the measure of dispersion and the acceptable margins of error to consider whether or not there exists a significant difference with respect to the average values or those that they are exactly within the regression line (95% of the data), or they are not outside the margins of the "Standard Error" (ES) which is expressed as "r" of Regression ("r" by Pearson)  $\pm$  ES

## MACHINE LEARNING WORKSHEET 1 (WORKSHEET SET 1)

The interpretation is that if most of the data are close to or around the regression line and its limits (standard error margins) that is, 95% of them; then, it can be said that the correlation and its estimation is significant.

In regression analysis, the distinction between errors and residuals is subtle and important, and leads to the concept of studentized residuals. Given an unobservable function that relates the independent variable to the dependent variable – say, a line – the deviations of the dependent variable observations from this function are the unobservable errors. If one runs a regression on some data, then the deviations of the dependent variable observations from the fitted function are the residuals.

However, a terminological difference arises in the expression mean squared error (MSE). The mean squared error of a regression is a number computed from the sum of squares of the computed residuals, and not of the unobservable errors. If that sum of squares is divided by  $n$ , the number of observations, the result is the mean of the squared residuals. Since this is a biased estimate of the variance of the unobserved errors, the bias is removed by dividing the sum of the squared residuals by  $df = n - p - 1$ , instead of  $n$ , where  $df$  is the number of degrees of freedom ( $n$  minus the number of parameters  $p$  being estimated - 1). This forms an unbiased estimate of the variance of the unobserved errors, and is called the mean squared error.[1]

Another method to calculate the mean square of error when analyzing the variance of linear regression using a technique like that used in ANOVA (they are the same because ANOVA is a type of regression), the sum of squares of the residuals (aka sum of squares of the error) is divided by the degrees of freedom (where the degrees of freedom equal  $n - p - 1$ , where  $p$  is the number of parameters estimated in the model (one for each variable in the regression equation)). One can then also calculate the mean square of the model by dividing the sum of squares of the model minus the degrees of freedom, which is just the number of parameters. Then the  $F$  value can be calculated by dividing the mean square of the model by the mean square of the error, and we can then determine significance.