MACHINE LEARNING

ASSIGNMENT – 6

1. In which of the following you can say that the model is overfitting?
Ans:
C) High R-squared value for train-set and Low R-squared value for test-set

2. Which among the following is a disadvantage of decision trees?
Ans:
B) Decision trees are highly prone to overfitting.

3.Which of the following is an ensemble technique?
Ans:
C) Random Forest

4.Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?
Ans:
C) Precision

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?
Ans:
B) Model B

6. Which of the following are the regularization technique in Linear Regression??
Ans:
A) Ridge

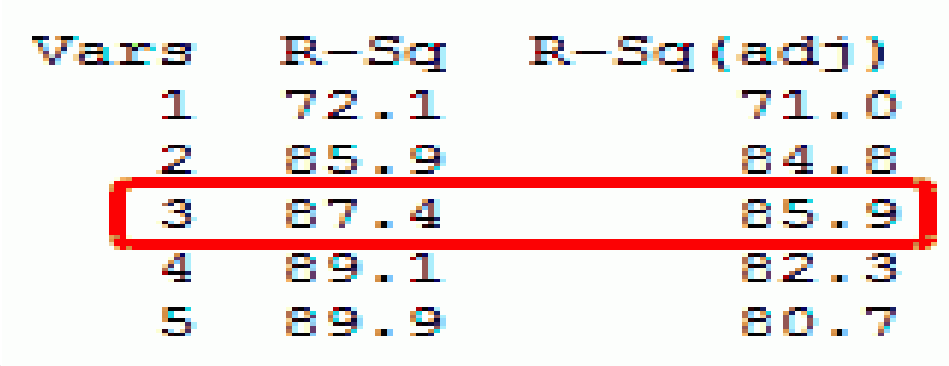
7. Which of the following is not an example of boosting technique?
Ans:
B) Decision Tree

8. Which of the techniques are used for regularization of Decision Trees?
Ans:
A) Pruning

9. Which of the following statements is true regarding the Adaboost technique?
Ans:
B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
C) It is example of bagging technique

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?
Ans:
Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than

expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.
When you are analyzing a situation in which there is a guarantee of little to no bias, using R-squared to calculate the relationship between two variables is perfectly useful. However, when investigating the relationship between say, the performance of a single stock and the rest of the S&P500, it is important to use adjusted R-squared to determine any inconsistencies in the correlation.



Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

.

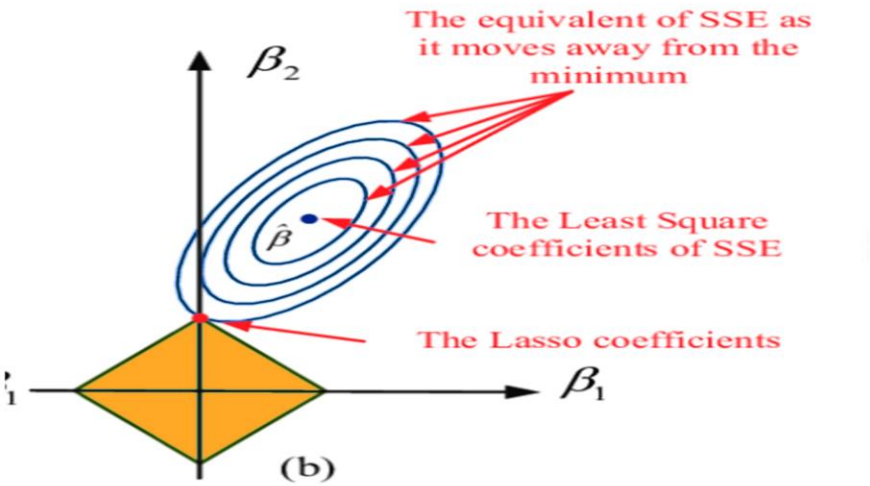
11.Differentiate between Ridge and Lasso Regression.

Lasso Regression

The word “LASSO” denotes Least Absolute Shrinkage and Selection Operator. Lasso regression follows the regularization technique to create prediction. It is given more priority over the other regression methods because it gives an accurate prediction. Lasso regression model uses shrinkage technique. In this technique, the data values are shrunk towards a central point similar to the concept of mean. The lasso regression algorithm suggests a simple, sparse models (i.e. models with fewer parameters), which is well-suited for models or data showing high levels of multicollinearity or when we would like to automate certain parts of model selection, like variable selection or parameter elimination using feature engineering.

Lasso Regression algorithm utilises L1 regularization technique It is taken into consideration when there are more number of features because it automatically performs feature selection.

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



Where,

- λ = the amount of shrinkage.
- If λ = 0 it implies that all the features are considered and now it is equivalent to the linear regression in which only the residual sum of squares is used to build a predictive model.
- If λ = ∞ it implies that no feature is used i.e, as λ gets close to infinity it eliminates more and more features and feature selection is more precise.
- When the bias increases, the value of λ increases
- When the variance increases, the value of λ decreases

Ridge Regression

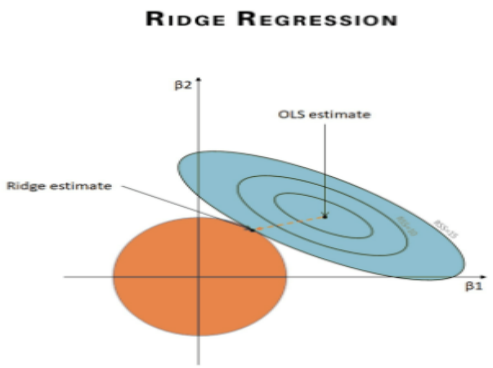
Ridge Regression is another type of regression algorithm in data science and is usually considered when there is a high correlation between the independent variables or model parameters. As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value. Therefore, we create a bias matrix in the equation of Ridge Regression algorithm. It is a useful regression method in which the model is less susceptible to overfitting and hence the model works well even if the dataset is very small.

First Pass

Digit	Number
0	620
1	891
2	902
3	243
4	134
5	655
6	426
7	
8	578
9	319

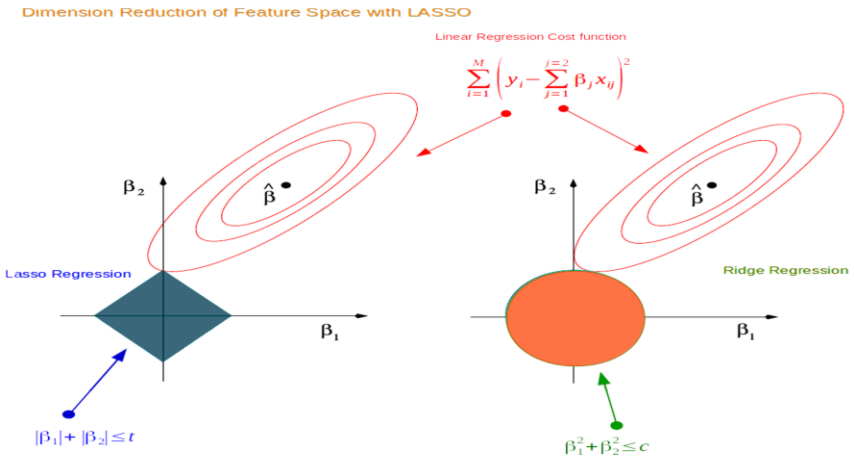
Where λ is the penalty variable. λ given here is denoted by an alpha parameter in the ridge function. Hence, by changing the values of alpha, we are controlling the penalty term. Greater the values of alpha, the higher is the penalty and therefore the magnitude of the coefficients is reduced.

We can conclude that it shrinks the parameters. Therefore, it is used to prevent multicollinearity, it also reduces the model complexity by shrinking the coefficient.



Comparision of both:

Ridge and Lasso regression uses two different penalty functions for regularisation. Ridge regression uses L2 on the other hand lasso regression go uses L1 regularisation technique. In ridge regression, the penalty is equal to the sum of the squares of the coefficients and in the Lasso, penalty is considered to be the sum of the absolute values of the coefficients. In lasso regression, it is the shrinkage towards zero using an absolute value (L1 penalty or regularization technique) rather than a sum of squares(L2 penalty or regularization technique).



12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans:
A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking an independent variable and regressing it against every other predictor in the model.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Including highly correlated variables in your model can lead to overfitting. If we overfit, then the model performs extraordinarily well on the training data but doesn't generalize well when we try to use it on new data.

Small VIF values, VIF < 3, indicate low correlation among variables under ideal conditions. The default VIF cutoff value is 5; only variables with a VIF less than 5 will be included in the model. However, note that many sources say that a VIF of less than 10 is acceptable.

Consider the following l regression model:

$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \varepsilon$
For each of the independent variables X_1, X_2 and X_3 we can calculate the variance inflation factor (VIF) in order to determine if we have a multicollinearity problem.
Here's the formula for calculating the VIF for X_1:

$$VIF_1 = \frac{1}{1 - R^2}$$

R^2 in this formula is the coefficient of determination from the regression model which has:

X_1 as dependent variable
X_2 and X_3 as independent variables
In other words, R^2 comes from the following linear regression model:
$X_1 = \beta_0 + \beta_1 \times X_2 + \beta_2 \times X_3 + \varepsilon$
And because R^2 is a number between 0 and 1:

When R^2 is close to 1 (i.e. X_2 and X_3 are highly predictive of X_1): the VIF will be very large
When R^2 is close to 0 (i.e. X_2 and X_3 are not related to X_1): the VIF will be close to 1

13.Why do we need to scale the data before feeding it to the train the model?
Ans:
To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model

I am not saying that all the algorithms will face this problem but most of the basic algorithms like linear and logistic regression, artificial neural networks, clustering algorithms with k value etc face the effect of the difference in scale for input variables

Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem. In the case of neural networks, an independent variable with a spread of values may result in a large loss training and testing and causing the learning process to be unstable.

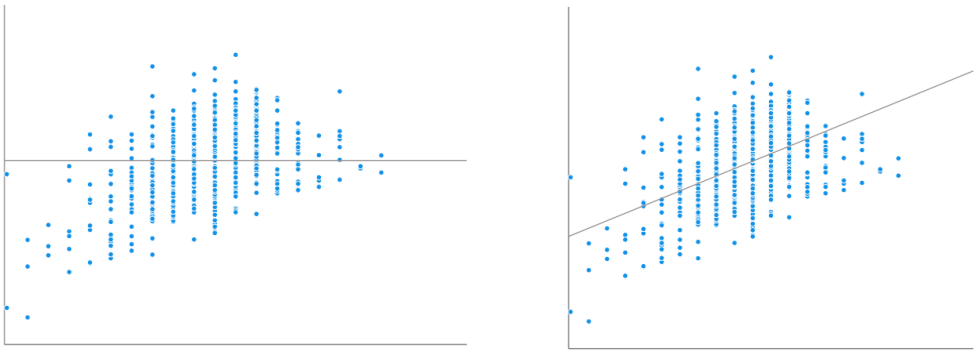
Normalization and Standardization are the two main methods for the scaling of the data. Which are widely used in the algorithms where scaling is required. Both Of them implemented by the scikit-learn libraries process package.

14.What are the different metrics which are used to check the goodness of fit in linear regression?

Ans:

A well-fitting regression model results in predicted values close to the observed data values.

The mean model, which uses the mean for every predicted value, generally would be used if there were no useful predictor variables. The fit of a proposed regression model should therefore be better than the fit of the mean model.



Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares Sum of Squares Total (SST) and Sum of Squares Error (SSE).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	640.816	1	640.816	560.782	<.001
Error	1368.977	1198	1.143		
Total	2009.793	1199			

SST measures how far the data are from the mean, and SSE measures how far the data are from the model's predicted values. Different combinations of these two values provide different information about how the regression model compares to the mean model.

R-squared:

The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the mode.

R-squared has the useful property that its scale is intuitive. It ranges from zero to one. Zero indicates that the proposed model does not improve prediction over the mean model. One indicates perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

One pitfall of R-squared is that it can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the model's fit. To remedy this, a related statistic, Adjusted R-squared, incorporates the model's degrees of freedom.

Adjusted R-squared:

Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Adjusted R-squared should always be used with models with more than one predictor variable. It is interpreted as the proportion of total variance that is explained by the model.

There are situations in which a high R-squared is not necessary or relevant. When the interest is in the relationship between variables, not in prediction, the R-squared is less important.

An example is a study on how religiosity affects health outcomes. A good result is a reliable relationship between religiosity and health. No one would expect that religion explains a high percentage of the variation in health, as health is affected by many other factors. Even if the model accounts for other variables known to affect health, such as income and age, an R-squared in the range of 0.10 to 0.15 is reasonable.

The F-test:

The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not. An equivalent null hypothesis is that R-squared equals zero.

A significant F-test indicates that the observed R-squared is reliable and is not a spurious result of oddities in the data set. Thus the F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable. It can be useful when the research objective is either prediction or explanation.

RMSE:

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance. It has the useful property of being in the same units as the response variable.

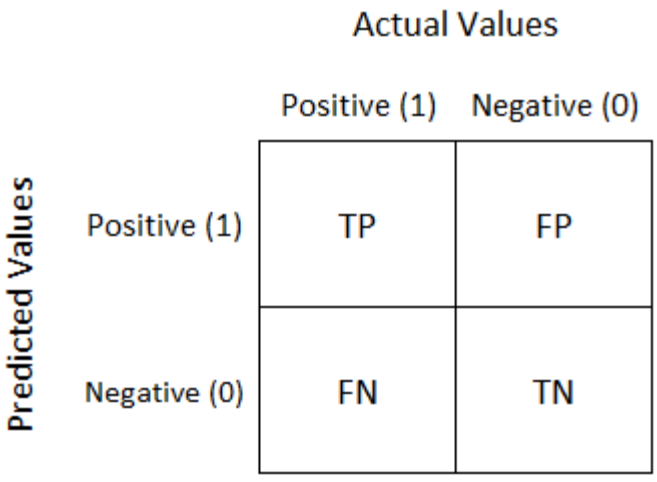
Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response. It's the most important criterion for fit if the main purpose of the model is prediction.

The best measure of model fit depends on the researcher's objectives, and more than one are often useful. The statistics discussed above are applicable to regression models that use OLS estimation.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans:



Confusion Metrics

From our confusion matrix, we can calculate five different metrics measuring the validity of our model.

Details-

TN = 1200
FN = 250
FP = 50
TP = 1000

Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN

1000+1200/2500=0.88 or 88% Accuracy

Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN

Precision (**true** positives / **predicted** positives) = TP / TP + FP

1000/1000+50=0.95 or 95% precision

Sensitivity aka Recall (**true** positives / all **actual** positives) = TP / TP + FN

1000/1000+250= 0.8 or 80% sensitivity

Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP

1200/1200+50=0.96 or 96% Specificity