MACHINE LEARNING ASSIGNMENT -39

Ans 1). (A) Least square error is used to find the best fit line for data in Linear Regression.

Ans 2). (A) Linear regression is sensitive to outliers. As outliers make the distribution somewhat skewed, so normal distribution condition is not satisfied in this case and hence, the accuracy of the model using Linear Regression is also decreased. To understand this, let us take the case that if there are outliers present in the linear regression model then the distance between these outliers and best fit line is very large as compared to other datapoints. So, will the sum of residuals (RSS) increase and as we know, rsquared value = (TSS -RSS)/TSS, so on increasing RSS, rsquared value decreases and hence, accuracy. Also, Mean Average Error (MAE) is directly related to the sum of squares of these residuals, so this error increases. That's the reason Linear Regression is sensitive to outliers.

Ans 3). (B) A line falls from left to right has NEGATIVE slope. As slope = (x2 - x1)/ (y2-y1) and if a line falls from left to right, it means x2>x1 and y2<y1 , therefore, x2-x1 is positive and y2-y1 is negative. Hence, the slope is negative.

Ans 4). (B) There is a symmetric relation between dependent variable and independent variable in case of Correlation only as correlation itself depicts that there is a strong and symmetric relationship between dependent variabkle and ind variable.

Ans 5). (C)Low bias and high variance are the reasons for Overfitting condition.

Ans 6). (B) If output involves label, then that model is called Predictive Model.

Ans 7). (D) Lasso and Ridge are the Regularization techniques. These will be discussed elaborately in answer no. 13.

Ans 8). (D) SMOTE is an upsampling technique used to overcome the problem of Imbalance dataset. Similarly, Near Miss is downsampling technique to handle imbalance dataset.

Ans 9). (A) AUC ROC curve is plotted between TPR vs FPR. TPR is senstivity and FPR is given by (1- Specificity).

Ans 10). (B) It is a False statement. In reality, for a better model, the area under the AUCROC curve should be more. More the area, better is the model or it can be interpreted generally as more TPR, more is the area and lower is the FPR, more is the area and better is the model.

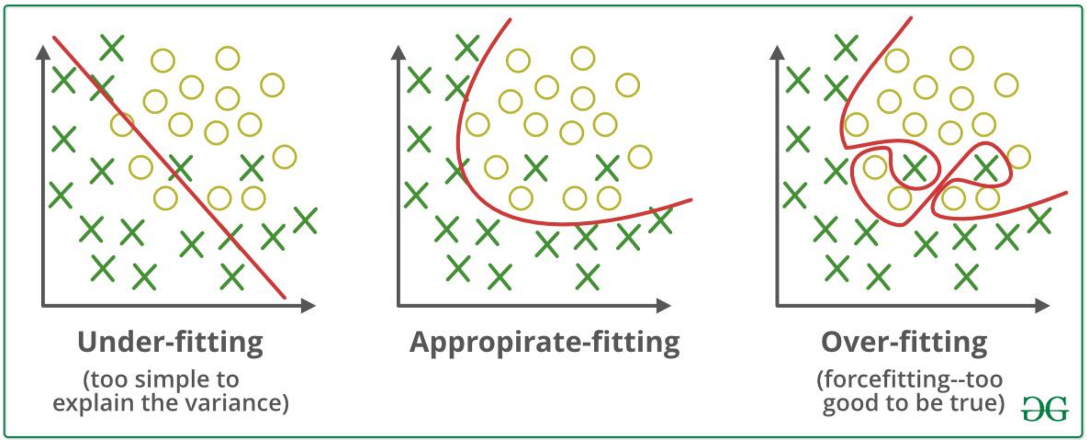
Ans 11). (B) Applying PCA to project high dimensional data is the right answer.

Ans 12). (A & B) are correct w.r.t. Normal equation. C is wrong as there is no need to iterate here. D is also wrong as Normal Equation uses dependent variable or target variable , y as clearly shown in its formula where x is an array of all independent features. T is the transpose of this array. And theta is Normal Equation,

$$\theta = \left(X^T X\right)^{-1} X^{\mathrm{T}} y$$

Ans 13) and Ans 14). **REGULARIZATION and Its Algorithms:**

Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data. So, to overcome this , there are some techniques which are called Regularization. These techniques reduce the errors by fitting the model on training set and thus avooid Overfitting.



The regularization is a form of regression that regularizes or shrinks the coefficients estimates towards zero. As we know that, the equation of linear regression is :

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

Where Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).
The coefficients are chosen such that the loss is minimized.

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

There are three regularization techniques. These are:
1) LASSO or L1 Form.
2) RIDGE or L2 Form.
3) ELASTICNET

1) LASSO Regularization or L1 form:

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.

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

It uses |βj|(modulus)instead of squares of β, as its penalty. 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 λ.
for lasso, the equation becomes,|β1|+|β2|≤ s. This implies that lasso coefficients have the smallest RSS(loss function) for all points that lie within the diamond given by |β1|+|β2|≤ s.
The L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
Its python implementation:

```
lasscv = LassoCV(alphas =None,max_iter = 100, normalize = True)
lasscv.fit(x_train, y_train)
lasscv.alpha_
```

The above command gives the most appropriate alpha value. Then putting this value, new model is formed which is free of overfitting problem.

```
lasso_model = Lasso(alpha = lasscv.alpha_)
Lasso_model.fit(x_train,y_train)
lasso_model.score(x_test,y_test)  which gives the improved accuracy than before.
```

2) RIDGE Regulaization or L2 form:

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

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

The above equation shows Ridge regularization basic principle. Here, RSS is modified by shrinking coefficients. λIs the tuning parameter. Now here, the increase in flexibility of model is governed by increase in its coefficients. So, if we want to minimize the above function, we need to lower down the coefficients as well. Thus, Ridge puts limit on increase of coefficients. So, if λ =0, then there is no effect. And if tends towards infinity, then the coefficients tend towards zero. So, selecting a critical value of λ is to be chosen.
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. 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.

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

In python, it is implemented as follows:

```
 ridge_cv  = RidgeCV (alphas = np.array(0.001,0.1,0.01), normalize = True)
ridge_cv.fit(x_train, y_train)
ridge_cv.alpha_
```

The above command gives the most appropriate alpha value. Then putting this value, new model is formed which is free of overfitting problem.

```
Ridge_model = Ridge(alpha = ridge_cv.alpha_)
Ridge_model.fit(x_train,y_train)
```

Ridge_model.score(x_test,y_test) which gives the improved accuracy than before.

3) **.ELATICNET Regularization** :

It’s a combination of both regularization techniques,I.e. it’sa combination of L1 and L2 form. It’s less popular than both of the above.

The elastic net method overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses a penalty function based on

$$\|\beta\|_1 = \textstyle\sum_{j=1}^{p} |\beta_j|.$$

Use of this penalty function has several limitations. For example, in the "large *p*, small *n*" case (high-dimensional data with few examples), the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty ($\|\beta\|^2$), which when used alone is ridge regression (known also as Tikhonov regularization). The estimates from the elastic net method are defined by

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Examples of where the elastic net method has been applied are:

- Support vector machine
- Metric learning
- Portfolio optimization
- Cancer prognosis

Ans 15). TERM ERROR PRESENT IN LINEAR REGRESSION EQUATION:

The error term used in the Linear Regression is the difference between the predicted value and the actual value which was to be predicted.An error term is a residual variable produced by 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. As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.

The error term is also known as the residual, disturbance, or remainder term, and is variously represented in models by the letters e, ε, or u.

In other words, it is the distance between the predicted value and the best fit line. Although the error term and residual are often used synonymously, there is an important formal difference. An error term is generally unobservable and a residual is observable and calculable, making it much easier to quantify and visualize. In effect, while an error term represents the way observed data differs from the actual population, a residual represents the way observed data differs from sample population data.

Error Term Use in a Formula

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a multiple linear regression function that takes the following form:

$Y=\alpha X+\beta \rho+\epsilon$ **where**

:α,β=Constant parameters

X,ρ=Independent variables

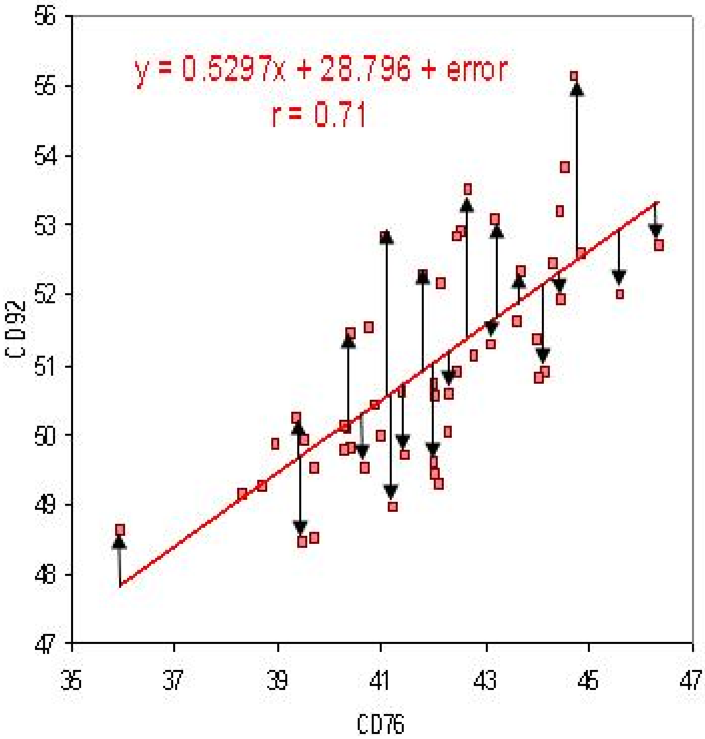
ϵ=Error term

Explained with example:

let’s say there is a study on the way the number of exams in a certain college affect the amount of red bull purchased from college vending machines. The data was collected which told that how many exams were given and how much red bull was purchased on a dozen or more days during the semester. This data can be plotted as a scatter plot, with exams (Ex) per given day on the x axis and red bull purchased (RB) per given day on the y axis. Then look at the line

$$y = \beta0 + \beta1x$$

that best fit the data.



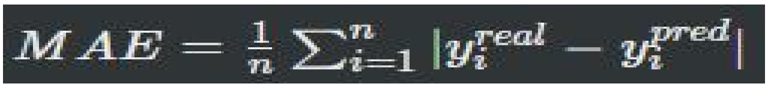
“Best fit” here means that the error term, the distance from each point to the line, is minimized. Since the relationship between variables is probably not completely linear and because there are other factors outside the scope of our study (sales on red bull, sales on other caffeine drinks, difficult physics homework sets, etc.) the graph won’t actually go through all our data points. 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

$$RB=\beta 0 + \beta 1 \text{ Ex} + \varepsilon$$

where β0 and β1 are constants and ε is an (non constant) error term.

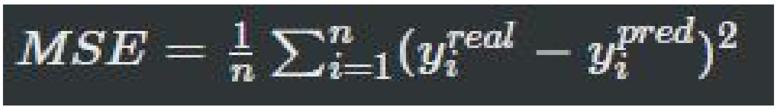
Usually, the metrics used to check accuracy or error presence in Linear Regression are of three types. These are:

1. Mean Absolute Error (MAE) : MAE evaluates the absolute distance of the observations (the entries of the dataset) to the predictions on a regression, taking the average over all observations. We use the absolute value of the distances so that negative errors are accounted properly. This is exactly the situation described on the image above.



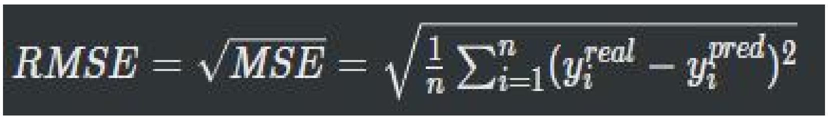
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i^{real} - y_i^{pred}|$$

2. Mean Square Error (MSE) : MSE evaluates the absolute distance of the observations (the entries of the dataset) to the predictions on a regression, taking the average over the squares of the residuals of all observations. Here, higher errors (or distances) weigh more in the metric than lower ones, due to the nature of the power function.



$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{real} - y_i^{pred})^2$$

3. Root Mean Square Error (RMSE): The disadvantage of MSE is the fact that the unit of the metric is also squared, so if the model tries to predict price in US$, the MSE will yield a number with unit (US$)² which does not make sense. RMSE is used then to return the MSE error to the original unit by taking the square root of it, while maintaining the property of penalizing higher errors.



$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{real} - y_i^{pred})^2}$$