Mock Takeaway Assignment

Name-Saurabh Sunil Kalal;BE(Comp Engg);April Batch

1. What are Eigenvectors and Eigenvalues? How are they relevant in the field of Data Science/Machine Learning?

Eigenvectors are column vectors or unit vectors whose length/magnitude is equal to 1. They are also called right vectors. Let A be an n×n matrix. A scalar λ is called an eigenvalue of A if the equation $Ax=\lambda x$ has a nonzero solution x. Such a nonzero solution x is called an eigenvector corresponding to the eigenvalue λ. Eigenvalues and Eigenvectors have a wide range of applications, for example in stability analysis, vibration analysis, atomic orbitals, facial recognition, and matrix diagonalization.

2. What do you understand about Imbalanced Data? How are these issues usually resolved?

Data is said to be highly imbalanced if it is distributed unequally across different categories. These datasets result in an error in model performance and result in inaccuracy.

There are different techniques to correct/balance imbalanced data. Following are some approaches followed to balance data:

1) Use the right evaluation metrics: (A) Precision/Specificity: how many selected instances are relevant. (B) Recall/Sensitivity: how many relevant instances are selected. (C)F1 score: harmonic mean of precision and recall. (D) MCC: correlation coefficient between the observed and predicted binary classifications. (F) AUC: relation between true-positive rate and false positive rate.

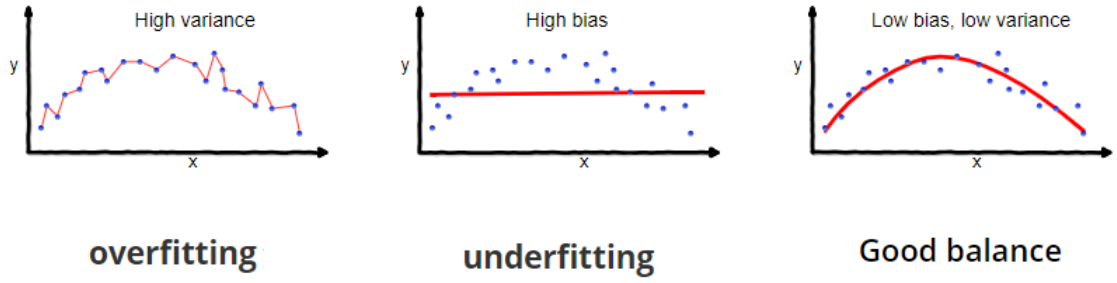
2)Training Set Resampling: Under-sampling & Over-sampling

3)Perform K-fold cross-validation correctly

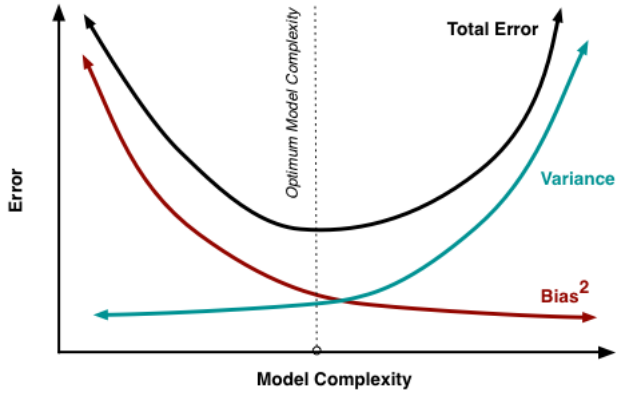
3.Define bias-variance trade-off?

Bias and Variance both are errors in machine learning models, it is very essential that any machine learning model has low variance as well as a low bias so that it can achieve good performance. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance

without overfitting and underfitting the data.



This tradeoff in complexity is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

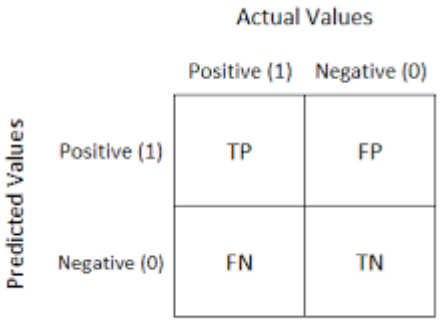


Total Error = Bias^2 + Variance + Irreducible Error

It helps optimize the error in our model and keeps it as low as possible.

4. Define the confusion matrix?

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. Confusion matrix is also termed as Error matrix.



True Positive(TP) [11] :- Both actual and predicted values are Positive.

True Negative(TN) [00] :- Both actual and predicted values are Negative.

False Positive(FP) [10] :- The actual value is negative but we predicted it as positive.

False Negative(FN) [01] :- The actual value is positive but we predicted it as negative.

CODE:

```
from sklearn import metrics
y_pred = ["a", "b", "c", "a", "b"] # Predicted values
y_act = ["a", "b", "c", "c", "a"] # Actual values
print(metrics.confusion_matrix(y_act, y_pred, labels=["a", "b", "c"]))# Printing the confusion matrix
# Printing the precision and recall, among other metrics
print(metrics.classification_report(y_act, y_pred, labels=["a", "b", "c"]))
```

5. What is Linear Regression? What are some of the major drawbacks of the linear model? State an example where you have recently used linear regression.

Linear regression comes under Regression part in Supervised Learning in Machining Learning.
It has both dependent and independent variable.
Linear regression is a technique in which the score of a variable Y is predicted using the score of a predictor variable X. Y is called the criterion variable.
A linear regression model attempts to explain the relationship between a dependent (output variables) variable and one or more independent (predictor variable) variables using a straight line.
This straight line is represented using the following formula:
$y = mx + c$
Where, y: dependent variable ; x: independent variable ; m: Slope of the line ; c: y intercept
It a statistical method that is used for predictive analysis.
There are two kinds of Linear Regression Model:-
Simple Linear Regression: A linear regression model with one independent and one dependent variable. Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable.

Some of the drawbacks of Linear Regression are as follows:

- ✓ The assumption of linearity of errors is a major drawback.
- ✓ It cannot be used for binary outcomes. We have Logistic Regression for that.
- ✓ Overfitting problems are there that can’t be solved.

Example :- Recently I had use linear regression in my project, where the csv file contains 2 features of Employee. In which salary column is dependent and year of experience is

independent variable. Linear regression a statistical method that is used for predictive analysis. As I was having both(salary & year of experience) the numeric variables, I choose to use Simple linear regression which makes predictions for continuous/real or numeric variables. Equation- Y=MX+C where Y is dependent variable(salary) & X is independent variable(year of experience).First I imported a file through pandas and check for null values, as there was no nan values then I slice the data and store the year of experience in X & salary in Y and convert it into numpy through .values in last of slicing part. Then I reshape the X & Y. Later I split into train and test part both the X & Y. And from sklearn.linear_model import LinearRegression. In Y_pred I stored the predicted values and Y_test the original values and then plotted the graph between those variables through matlotlib.pyplot module. Where I get a straight line (predicted line).

6. What is a Gradient and Gradient Descent? How is it used? Show with an example.

Gradient: Gradient is the measure of a property that how much the output has changed with respect to a little change in the input. In other words, we can say that it is a measure of change in the weights with respect to the change in error. The gradient can be mathematically represented as the slope of a function.

Gradient descent: Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. Linear Regression and Logistic Regression use gradient descent. It is one of the most used machine learning algorithms in the industry

Example: First we have to perform the linear regression, as I mentioned in the above question But there can be multiple lines that can pass through these points. So first we have to perform cost function.

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2$$

Replace $y_{i\,pred}$ with $mx_i + c$

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

I had not try all the permutation and combination of m and c (inefficient way) to find the best-fit line. For that,I had use Gradient Descent Algorithm.

m = 0 and c = 0. L be our learning rate. It could be a small value like 0.01 for good accuracy. Put it to zero means our model isn't learning anything from the gradients. Later I Calculated the partial derivative of the Cost function with respect to m. Here I had taken Dm as partial derivative of the Cost function with respect to m

$$D_m = \frac{\partial(Cost\ Function)}{\partial m} = \frac{\partial}{\partial m}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial m}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - (mx_i + c))$$

$$= \frac{-2}{n}\sum_{i=0}^{n}x_i(y_i - y_{i\,pred})$$

$$D_c = \frac{\partial(Cost\ Function)}{\partial c} = \frac{\partial}{\partial c}\left(\frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i - (mx_i + c))^2\right)$$

$$= \frac{1}{n}\frac{\partial}{\partial c}\left(\sum_{i=0}^{n}(y_i^2 + m^2x_i^2 + c^2 + 2mx_ic - 2y_imx_i - 2y_ic)\right)$$

$$= \frac{-2}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))$$

$$\frac{-2}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})$$

3. Now update the current values of m and c using the following equation:

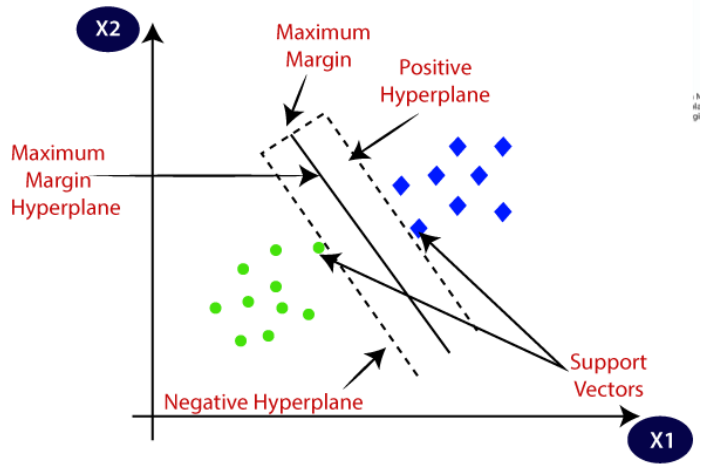
m=m-LDc & c=c-LDc.

4. I had repeated this process until the Cost function is very small (ideally 0).

Gradient Descent Algorithm gives optimum values of m and c of the linear regression equation. With these values of m and c, we will get the equation of the best-fit line and ready to make predictions.

7. What are Support Vectors in SVM (Support Vector Machine)? Explain with a use case.

Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.



Support Vector Machine(SVM) comes under Classification part in Supervised Learning in Machine Learning. In above diagram the middle dark line is kernel which separate two classes.

CODE: from sklearn.svm import svc #importing svc from Scikit-learn
model=svc(kernel='linear',random_state=0)
model.fit(X_train,Y_train) #here we train our model through X_train & Y_train variables

First I read our file, check for null values, if yes perform OneHotEncoder and column transformer on that and then reshape the variables and perform standard scaler on it. Splited the variables in train and test, and then perform the above code i.e svc and stored the predicted value in Y_pred. then perform confusion matrix and accuracy score. Later get the accuracy score by Y_pred & Y_test. And later change the kernel from linear to poly and degree to 2 and check for accuracy score. And also make kernel=rbf and check the score. I come to know that it increases the accuracy score.

8. How would you handle a dataset with missing values of more than 30%?

Checking null values:- 1)In pandas df[‘column_name’].isnull().any() -> gives the Boolean value false(not null values) or True(null values). 2)And data.isnull().sum() -> gives the total sum of null values of a particular columns where data is a csv file or any data.

Depending on the size of the dataset, we follow the below ways:

1)In case the datasets are small, the missing values are substituted with the mean or average of the remaining data. In pandas, this can be done by using mean = df.mean() where df represents the pandas dataframe representing the dataset and mean() calculates the mean of the data. To substitute the missing values with the calculated mean, we can use df.fillna(mean).

CODE(filling nan values by SimpleImputer):

```
from sklearn.impute import SimpleImputer       # Importing the SimpleImputer class

# Imputer object using the mean strategy

si = SimpleImputer(missing_values = np.nan, strategy ='mean')

data = si.fit_transform(data)   # Imputing the data
```

CODE(filling nan values by .fillna()):

```
df[‘column_name’]= df[‘column_name’].fillna(df[‘column_name’].mean()) #by mean

df[‘column_name’]= df[‘column_name’].fillna(df[‘column_name’].median()) #by median
```

CODE(deleting the columns):

```
df.columns #it will give the columns name

reduced_df=df.drop(cols_with_missing, axis=1)
```

where axis=0 means rows ; axis=1 means columns

2)For larger datasets, the rows with missing values can be removed and the remaining data can be used for data prediction.

CODE(deleting the rows):

```
df.dropna([5,6], axis=0, inplace=True)

print(df)

#here,  [5,6] is the index of the rows you want to delete

axis=0 denotes that rows should be deleted from the dataframe

inplace=True performs the drop operation in the same dataframe
```

CODE(Dropping Rows with at least 1 null value in CSV file)

```
new_data = data.dropna(axis = 0, how ='any')

    print(new_data)
```

9. Why is data cleaning crucial? How do you clean the data?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. While running an algorithm on any data, to gather proper insights, it is very much necessary to have correct and clean data that contains only relevant information. Dirty data most often results in poor or incorrect insights and predictions which can have damaging effects.

For example, while launching any big campaign to market a product, if our data analysis tells us to target a product that in reality has no demand and if the campaign is launched, it is bound to fail. This results in a loss of the company's revenue. This is where the importance of having proper and clean data comes into the picture.

Data Cleaning of the data coming from different sources helps in data transformation and results in the data where the data scientists can work on.

Properly cleaned data increases the accuracy of the model and provides very good predictions.

Data cleaning helps to identify and fix any structural issues in the data. It also helps in removing any duplicates and helps to maintain the consistency of the data.

Data Cleaning is a process which comes under Data Preprocessing process. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various

ways.
Some of them are:

- Ignore the tuples:
  This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

- Fill the Missing values:
  There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:
Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

Binning Method:
This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Regression:
Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:
This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

10. An in-depth analysis of an organization's ideal customers is referred to as a Customer Personality Analysis. Businesses can use it to better understand their customers and modify products according to their preferences, behavior, and concerns. Suppose you have to do an analysis that should help a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment. Explain in detail your approach how you would solve this problem for the company - beginning from the data collection to model selection to deployment.

Step-1: Problem Characterization: The first step I will take is to thoroughly understand the business requirement/problem and understanding the target requirements. Also I will clearly identify different relevant independent variables and dependent variables.

Step-2: Data Collection: In second step I will collect relevant and comprehensive data. May be I will collect real time data depending on the type of data analytics. As no one has infinite resources and infinite time to collect fully comprehensive data, most relevant representative data should be collected.

Step-3: Data Preparation: Data preparation is crucial process that deals with preparing the data for the model development. Next, I will explore the given data and analyze it carefully. I will perform data cleansing, labeling the data, dealing with missing data, OneHotEncoder and dealing with inconsistent data. Filling missing values with mean, ,median of column, or deleting the rows having nan values according to relevant approach. KNN (k-nearest neighbors): Fill data with a value from another example that is similar. And then performing LabelEncoding(mapping string with a numbers) and OneHotEnding(0's and 1's) with columns.

Data splitting: Training set (usually 70-80% of data): Model learns on this. Test set (usually 10-15% of data): Model's final performance is evaluated on this. Train model on data(3 steps: Choose an algorithm, overfit the model, reduce overfitting with regularization)

Choosing an algorithms:(A) Supervised algorithms – Linear Regression, Logistic Regression, KNN, SVMs, Decision tree and Random forests, AdaBoost/Gradient Boosting Machine(boosting). (B) Unsupervised algorithms- Clustering, dimensionality reduction(PCA, Autoencoders, t-SNE), An anomaly detection.

Then I will check for Regression and Classification Approach.

Step-4: Data Exploration: here I will go through Exploratory Data Analysis(EDA) and extract usefull insights. Next I will use Business Intelligence tools like Tableau, PowerBi which can be quite beneficial in this step.

Step-5: Modeling: During the modeling step, ML model is selected, trained, validated and tested. The common approach to build a good model is try to different algorithms and compare their performance. I will run the model against the data, and build meaningful visualization and analyse the results to get meaningful insights.

Step-6: Evaluation: This is the next phase, it is crucial to check that our Data Science Modelling efforts meet the expectations. The Data Model is applied to the Test Data to check if it's accurate. I will further test the Data Model to identify any adjustments that might be required to enhance the performance and achieve the desired results. If the required precision is not achieved, I wii go back to Step 5 (Machine Learning Algorithms), choose an alternate Data Model, and then test the model again.

Step-7: Model Deployment: Once the model is developed and optimized, it can be deployed into the system/process. I will check for the Model which provides the best result based on test findings is completed and deployed in the production environment whenever the

desired result is achieved through proper testing as per the business needs. This concludes the process of Data Science Modelling