**Interview Question on Machine Learning**

**Naive Bayes Classifier:-**

##### *1. What Are the Basic Assumption?*

Features Are Independent

##### *2. Advantages*

1. Work Very well with many number of features
2. Works Well with Large training Dataset
3. It converges faster when we are training the model
4. It also performs well with categorical features

##### *3. Disadvantages*

1. Correlated features affects performance

##### *4. Whether Feature Scaling is required?*

No

##### *5. Impact of Missing Values?*

Naive Bayes can handle missing data. Attributes are handled separately by the algorithm at both model construction time and prediction time. As such, if a data instance has a missing value for an attribute, it can be ignored while preparing the model, and ignored when a probability is calculated for a class value tutorial :<https://www.youtube.com/watch?v=EqjyLfpv5oA>

##### *6. Impact of outliers?*

It is usually robust to outliers

##### *Different Problem statement you can solve using Naive Baye's*

1. Sentiment Analysis
2. Spam classification
3. twitter sentiment analysis
4. document categorization

### How To Learn Machine Learning Algorithms For Interviews

#### Linear Regression

Theoretical Understanding:

1. <https://www.youtube.com/watch?v=1-OGRohmH2s&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=29>
2. <https://www.youtube.com/watch?v=5rvnlZWzox8&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=34>
3. <https://www.youtube.com/watch?v=NAPhUDjgG_s&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=32>
4. <https://www.youtube.com/watch?v=WuuyD3Yr-js&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=35>
5. <https://www.youtube.com/watch?v=BqzgUnrNhFM&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=33>

##### *Interview Question on Multicollinearity*

1. <https://www.youtube.com/watch?v=tcaruVHXZwE>

##### *1. What Are the Basic Assumption?(favourite)*

There are four assumptions associated with a linear regression model:

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

##### *2. Advantages*

1. Linear regression performs exceptionally well for linearly separable data
2. Easy to implement and train the model
3. It can handle overfitting using dimensionlity reduction techniques and cross validation and regularization

##### *3. Disadvantages*

1. Sometimes Lot of Feature Engineering Is required
2. If the independent features are correlated it may affect performance
3. It is often quite prone to noise and overfitting

##### *4. Whether Feature Scaling is required?*

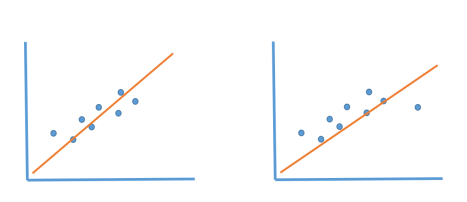
Yes

##### *5. Impact of Missing Values?*

It is sensitive to missing values

##### *6. Impact of outliers?*

linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects.



##### *Types of Problems it can solve(Supervised)*

1. Regression

##### *Overfitting And Underfitting*

HomeWork?

##### *Different Problem statement you can solve using Linear Regression*

1. Advance House Price Prediction
2. Flight Price Prediction

#### Practical Implementation

1. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

### How To Learn Machine Learning Algorithms For Interviews

#### SVM

Theoretical Understanding:

1. <https://www.youtube.com/watch?v=H9yACitf-KM>
2. <https://www.youtube.com/watch?v=Js3GLb1xPhc>

##### *1. What Are the Basic Assumption?*

There are no such assumptions

##### *2. Advantages*

1. SVM is more effective in high dimensional spaces.
2. SVM is relatively memory efficient.
3. SVM’s are very good when we have no idea on the data.
4. Works well with even unstructured and semi structured data like text, Images and trees.
5. The kernel trick is real strength of SVM. With an appropriate kernel function, we can solve any complex problem.
6. SVM models have generalization in practice, the risk of over-fitting is less in SVM.

##### *3. Disadvantages*

1. More Training Time is required for larger dataset
2. It is difficult to choose a good kernel function <https://www.youtube.com/watch?v=mTyT-oHoivA>
3. The SVM hyper parameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact

##### *4. Whether Feature Scaling is required?*

Yes

##### *5. Impact of Missing Values?*

Although SVMs are an attractive option when constructing a classifier, SVMs do not easily accommodate missing covariate information. Similar to other prediction and classification methods, in-attention to missing data when constructing an SVM can impact the accuracy and utility of the resulting classifier.

##### *6. Impact of outliers?*

It is usually sensitive to outliers <https://arxiv.org/abs/1409.0934#:~:text=Despite%20its%20popularity%2C%20SVM%20has,causes%20the%20sensitivity%20to%20outliers>.

##### *Types of Problems it can solve(Supervised)*

1. Classification
2. Regression

##### *Overfitting And Underfitting*

In SVM, to avoid overfitting, we choose a Soft Margin, instead of a Hard one i.e. we let some data points enter our margin intentionally (but we still penalize it) so that our classifier don't overfit on our training sample

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

##### *Different Problem statement you can solve using Naive Baye's*

1. We can use SVM with every ANN usecases
2. Intrusion Detection
3. Handwriting Recognition

#### Practical Implementation

1. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

##### *Performance Metrics*

##### *Classification*

1. Confusion Matrix
2. Precision,Recall, F1 score

##### *Regression*

1. R2,Adjusted R2
2. MSE,RMSE,MAE

### How To Learn Machine Learning Algorithms For Interviews

#### Decision Tree Classifier And Regressor

Interview Questions:

1. Decision Tree
2. Entropy, Information Gain, Gini Impurity
3. Decision Tree Working For Categorical and Numerical Features
4. What are the scenarios where Decision Tree works well
5. Decision Tree Low Bias And High Variance- Overfitting
6. Hyperparameter Techniques
7. Library used for constructing decision tree
8. Impact of Outliers Of Decision Tree
9. Impact of mising values on Decision Tree
10. Does Decision Tree require Feature Scaling

Theoretical Understanding:

1. Tutorial 37:Entropy In Decision Tree <https://www.youtube.com/watch?v=1IQOtJ4NI_0>
2. Tutorial 38:Information Gain <https://www.youtube.com/watch?v=FuTRucXB9rA>
3. Tutorial 39:Gini Impurity <https://www.youtube.com/watch?v=5aIFgrrTqOw>
4. Tutorial 40: Decision Tree For Numerical Features: <https://www.youtube.com/watch?v=5O8HvA9pMew>
5. How To Visualize DT: <https://www.youtube.com/watch?v=ot75kOmpYjI>

##### *1. What Are the Basic Assumption?*

There are no such assumptions

##### *2. Advantages*

Advantages of Decision Tree

1. **Clear Visualization: The algorithm is simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.**
2. Simple and easy to understand: Decision Tree looks like simple if-else statements which are very easy to understand.
3. Decision Tree can be used for both classification and regression problems.
4. Decision Tree can handle both continuous and categorical variables.
5. No feature scaling required: No feature scaling (standardization and normalization) required in case of Decision Tree as it uses rule based approach instead of distance calculation.
6. Handles non-linear parameters efficiently: Non linear parameters don't affect the performance of a Decision Tree unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Decision Trees may outperform as compared to other curve based algorithms.
7. Decision Tree can automatically handle missing values.
8. Decision Tree is usually robust to outliers and can handle them automatically.
9. Less Training Period: Training period is less as compared to Random Forest because it generates only one tree unlike forest of trees in the Random Forest.

##### *3. Disadvantages*

Disadvantages of Decision Tree

1. Overfitting: This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.
2. High variance: As mentioned in point 1, Decision Tree generally leads to the overfitting of data. Due to the overfitting, there are very high chances of high variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results. In order to achieve zero bias (overfitting), it leads to high variance.
3. Unstable: Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.
4. Not suitable for large datasets: If data size is large, then one single tree may grow complex and lead to overfitting. So in this case, we should use Random Forest instead of a single Decision Tree.

##### *4. Whether Feature Scaling is required?*

No

##### *6. Impact of outliers?*

It is not sensitive to outliers.Since, extreme values or outliers, never cause much reduction in RSS, they are never involved in split. Hence, tree based methods are insensitive to outliers.

##### *Types of Problems it can solve(Supervised)*

1. Classification
2. Regression

##### *Overfitting And Underfitting*

Ho to avoid overfitting

<https://www.youtube.com/watch?v=SLOyyFHbiqo>

#### Practical Implementation

1. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>

##### *Performance Metrics*

##### *Classification*

1. Confusion Matrix
2. Precision,Recall, F1 score

##### *Regression*[*¶*](https://render.githubusercontent.com/view/ipynb?color_mode=light&commit=d849ef8f951e217fc8e1145d3ba2392e828610c5&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f6b726973686e61696b30362f496e746572766965772d50726570617274696f6e2d446174612d536369656e63652f643834396566386639353165323137666338653131343564336261323339326538323836313063352f496e746572766965772532305072657061726174696f6e2d253230446179253230342d2532304465636973696f6e25323054726565732e6970796e62&nwo=krishnaik06%2FInterview-Prepartion-Data-Science&path=Interview+Preparation-+Day+4-+Decision+Trees.ipynb&repository_id=297263348&repository_type=Repository#Regression)

1. R2,Adjusted R2
2. MSE,RMSE,MAE

### How To Learn Machine Learning Algorithms For Interviews

#### Logistics Regression

Theoretical Understanding:

1. Tutorial 35:Logitic Regression Part 1 <https://www.youtube.com/watch?v=L_xBe7MbPwk>
2. Tutorial 36:Logitic Regression Part 2 <https://www.youtube.com/watch?v=uFfsSgQgerw>
3. Tutorial 39:Logitic Regression Part 3 <https://www.youtube.com/watch?v=V8fS0T_ktn4>
4. Tutorial 42:How To Find Optimal Threshold for Binary classification: <https://www.youtube.com/watch?v=_AjhdXuXEDE>
5. Interview question: <https://www.youtube.com/watch?v=tcaruVHXZwE&t=122s>

##### *1. What Are the Basic Assumption?*

1. Linear Relation between independent features and the log odds

##### *2. Advantages*

Advantages of Logistics Regression

1. Logistic Regression Are very easy to understand
2. It requires less training
3. Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
4. It makes no assumptions about distributions of classes in feature space.
5. Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.One may consider Regularization (L1 and L2) techniques to avoid over-fittingin these scenarios.
6. Logistic regression is easier to implement, interpret, and very efficient to train.

##### *3. Disadvantages*

1. Sometimes Lot of Feature Engineering Is required
2. If the independent features are correlated it may affect performance
3. It is often quite prone to noise and overfitting
4. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
5. Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
6. It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.
7. In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds (log(p/(1-p)).

##### *4. Whether Feature Scaling is required?*

yes

#### 5. Missing Values

Sensitive to missing values

##### *6. Impact of outliers?*

Like linear regression, estimates of the logistic regression are sensitive to the unusual observations: outliers, high leverage, and influential observations. Numerical examples and analysis are presented to demonstrate the most recent outlier diagnostic methods using data sets from medical domain

##### *Types of Problems it can solve(Supervised)*

1. Classification

#### Practical Implementation

1. <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

##### *Performance Metrics*

##### *Classification*

1. Confusion Matrix
2. Precision,Recall, F1 score
3. Part 1 <https://www.youtube.com/watch?v=aWAnNHXIKww>
4. Part 2 <https://www.youtube.com/watch?v=A_ZKMsZ3f3o>

### How To Learn Machine Learning Algorithms For Interviews

#### Decision Tree Classifier And Regressor

Interview Questions:

1. Decision Tree
2. Entropy, Information Gain, Gini Impurity
3. Decision Tree Working For Categorical and Numerical Features
4. What are the scenarios where Decision Tree works well
5. Decision Tree Low Bias And High Variance- Overfitting
6. Hyperparameter Techniques
7. Library used for constructing decision tree
8. Impact of Outliers Of Decision Tree
9. Impact of mising values on Decision Tree
10. Does Decision Tree require Feature Scaling

##### *Random Forest Classifier And Regresor*

1. Ensemble Techniques(Boosting And Bagging)
2. Working of Random Forest Classifier
3. Working of Random Forest Regresor
4. Hyperparameter Tuning(Grid Search And RandomSearch)

Theoretical Understanding:

1. Tutorial 37:Entropy In Decision Tree <https://www.youtube.com/watch?v=1IQOtJ4NI_0>
2. Tutorial 38:Information Gain <https://www.youtube.com/watch?v=FuTRucXB9rA>
3. Tutorial 39:Gini Impurity <https://www.youtube.com/watch?v=5aIFgrrTqOw>
4. Tutorial 40: Decision Tree For Numerical Features: <https://www.youtube.com/watch?v=5O8HvA9pMew>
5. How To Visualize DT: <https://www.youtube.com/watch?v=ot75kOmpYjI>

Theoretical Understanding:

1. Ensemble technique(Bagging): <https://www.youtube.com/watch?v=KIOeZ5cFZ50>
2. Random forest Classifier And Regressor <https://www.youtube.com/watch?v=nxFG5xdpDto>
3. Construct Decision Tree And working in Random Forest: <https://www.youtube.com/watch?v=WQ0iJSbnnZA&t=406s>

#### Important properties of Random Forest Classifiers

1. Decision Tree---Low Bias And High Variance
2. Ensemble Bagging(Random Forest Classifier)--Low Bias And Low Variance

##### *1. What Are the Basic Assumption?*

There are no such assumptions

##### *2. Advantages*

Advantages of Random Forest

1. **Doesn't Overfit**
2. Favourite algorithm for Kaggle competition
3. Less Parameter Tuning required
4. Decision Tree can handle both continuous and categorical variables.
5. No feature scaling required: No feature scaling (standardization and normalization) required in case of Random Forest as it uses DEcision Tree internally
6. Suitable for any kind of ML problems

##### *3. Disadvantages*

Disadvantages of Random Forest

1.**Biased With features having many categories**

1. Biased in multiclass classification problems towards more frequent classes.

##### *4. Whether Feature Scaling is required?*

No

##### *6. Impact of outliers?*

Robust to Outliers

##### *Types of Problems it can solve(Supervised)*

1. Classification
2. Regression

#### Practical Implementation

1. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
2. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
3. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
4. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

##### *Performance Metrics*

##### *Classification*

1. Confusion Matrix
2. Precision,Recall, F1 score

##### *Regression*

1. R2,Adjusted R2
2. MSE,RMSE,MAE

### How To Learn Machine Learning Algorithms For Interviews

#### Decision Tree Classifier And Regressor

Interview Questions:

1. Decision Tree
2. Entropy, Information Gain, Gini Impurity
3. Decision Tree Working For Categorical and Numerical Features
4. What are the scenarios where Decision Tree works well
5. Decision Tree Low Bias And High Variance- Overfitting
6. Hyperparameter Techniques
7. Library used for constructing decision tree
8. Impact of Outliers Of Decision Tree
9. Impact of mising values on Decision Tree
10. Does Decision Tree require Feature Scaling

#### Xgboost Classifier And Regressor, GB Algorithm, Adaboost

Decision Tree Theoretical Understanding:

1. Tutorial 37:Entropy In Decision Tree <https://www.youtube.com/watch?v=1IQOtJ4NI_0>
2. Tutorial 38:Information Gain <https://www.youtube.com/watch?v=FuTRucXB9rA>
3. Tutorial 39:Gini Impurity <https://www.youtube.com/watch?v=5aIFgrrTqOw>
4. Tutorial 40: Decision Tree For Numerical Features: <https://www.youtube.com/watch?v=5O8HvA9pMew>
5. How To Visualize DT: <https://www.youtube.com/watch?v=ot75kOmpYjI>

Theoretical Understanding:

1. Ensemble technique(Bagging): <https://www.youtube.com/watch?v=KIOeZ5cFZ50>
2. Adaboost(Boosting Technique):<https://www.youtube.com/watch?v=NLRO1-jp5F8>
3. Gradient Boosting In Depth Intuition Part 1: <https://www.youtube.com/watch?v=Nol1hVtLOSg>
4. Gradient Boosting In Depth Intuition Part 2: <https://www.youtube.com/watch?v=Oo9q6YtGzvc>
5. Xgboost Classifier Indepth Intuition: <https://www.youtube.com/watch?v=gPciUPwWJQQ>
6. Xgboost Regression Indpeth Intuition: <https://www.youtube.com/watch?v=w-_vmVfpssg>
7. Implementation of Xgboost: <https://youtu.be/9HomdnM12o4>

##### *1. What Are the Basic Assumption?*

There are no such assumptions

#### Missing Values

1. Adaboost can handle mising values
2. Xgboosst and GBoost cannot handle missing values

##### *2. Advantages*

Advantages of Adaboost

1. **Doesn't Overfit**
2. It has few parameters to tune

Advantages of Gradient Boost And Xgboost

1. It has a great performance
2. It can solve complex non linear functions
3. It is better in solve any kind of ML usecases.

##### *3. Disadvantages*

Disadvantages of Gradient Boosting And Xgboost

1.It requires some amount of parameter tuning

##### *4. Whether Feature Scaling is required?*

No

##### *6. Impact of outliers?*

Robust to Outliers in Gradient Boosting And Xgboost, Sensitive to outliers in Adaboost

##### *Types of Problems it can solve(Supervised)*

1. Classification
2. Regression

##### *Performance Metrics*

##### *Classification*

1. Confusion Matrix
2. Precision,Recall, F1 score

##### *Regression*

1. R2,Adjusted R2
2. MSE,RMSE,MAE