# Introduction to Machine Learning

1. What is the differences between supervised machine learning and unsupervised machine learning?

A:

Supervised Learning: if you are training your machine learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training.

Unsupervised Learning: if you are training your machine learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs.

1. What is the differences between classification and regression?

A:

Classification: Identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

Regression: assuming the target values come from a sort of function and focused on finding the relationship between input variables and the continuous outputs.

1. Explain the concept of machine learning in very simple words.

# Linear Regression

1. What’s the assumption of Linear Regression?

A:

* The relationship between the independent variables and dependent variables should be linear relationship.
* Among the features there is little or no linear correlations.
* All the features have a normal distribution
* No auto-correlation
* The error term along the regression are equal

1. What’s the disadvantages of linear regression?

A:

* Very sensitive to outliers
* It can only describe linear relationship between the variables.
* It requires a lot of work on data preprocessing because of so many assumptions.

1. For the univariate problem, which are the ordinary least squares solution for linear regression?

A:





1. Which is the general OLS solution for linear regression in the following:

A:



Logistic Regression

1. What’s the differences between logistic regression and linear regression?

A:

* Linear regression output can be out of range 0-1 while logistic regression cannot
* Linear regression aimed to solve regression while logistic aimed to solve classification
* Logistic Regression is used when response variable is categorical in nature. Linear Regression is used when your response variable is continuous.
* Linear Regression gives an equation which is of the form Y = mX + C, means equation with degree 1. However, Logistic Regression gives an equation which is of the form Y = e^X/1 + e^-X (there is a logistic function)
* Coefficient interpretation

In linear regression, the coefficient interpretation of independent variables is quite straight forward. However, in logistic regression, depends on the family (binomial, poisson, etc.) and link (log, logit, inverse-log, etc.) you use, the interpretation is different.

* Error Minimization Technique

Linear Regression uses Ordinary Least Squares method to minimize the errors and arrive at a best possible fit. While logistic regression uses maximum likelihood method to arrive at the solution.

Decision Tree & Random Forest

1. Describe what is a decision tree.

A:

Decision tree is a type of supervised learning algorithm. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

1. What’s the advantages of decision tree?

A:

Easy to Understand: Decision tree output is very easy to understand. Its graphical representation is very intuitive and users can easily relate their hypothesis.

Useful in Data exploration: Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.

Less data cleaning required: It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.

Data type is not a constraint: It can handle both numerical and categorical variables.

Non Parametric Method: Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

1. What’s the disadvantages of decision tree?

A:  
Overfitting: Overfitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning.

Not fit for continuous variables: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories

1. How does a tree decide where to split.

A:

The decision criteria is different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. Some most commonly used algorithms to split the node are: Gini Index, Chi-Square, Information Gain, Reduction in Variance.

Third Party Questions:

**1) True-False: Linear Regression is a supervised machine learning algorithm.**

A) TRUE  
B) FALSE

**Solution: (A)**

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

**2) True-False: Linear Regression is mainly used for Regression.**

A) TRUE  
B) FALSE

**Solution: (A)**

**Linear Regression** has dependent variables that have continuous values.

**3) True-False: It is possible to design a Linear regression algorithm using a neural network?**

A) TRUE  
B) FALSE

**Solution: (A)**

True. A Neural network can be used as a *universal* approximator, so it can definitely implement a linear regression algorithm.

**4) 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

**Solution: (A)**

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

**5) Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?**

A) AUC-ROC  
B) Accuracy  
C) Logloss  
D) Mean-Squared-Error

**Solution: (D)**

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

**6) True-False: Lasso Regularization can be used for variable selection in Linear Regression.**

A) TRUE  
B) FALSE

**Solution: (A)**

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

**7) Which of the following is true about Residuals ?**

A) Lower is better  
B) Higher is better  
C) A or B depend on the situation  
D) None of these

**Solution: (A)**

Residuals refer to the error values of the model. Therefore lower residuals are desired.

**8)**Suppose that we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of it’s variable(Say X1) with Y is -0.95.

**Which of the following is true for X1?**

A) Relation between the X1 and Y is weak  
B) Relation between the X1 and Y is strong  
C) Relation between the X1 and Y is neutral  
D) Correlation can’t judge the relationship

**Solution: (B)**

The absolute value of the correlation coefficient denotes the strength of the relationship. Since  absolute correlation is very high it means that the relationship is strong between X1 and Y.

**9) Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?**

If you are given the two variables V1 and V2 and they are following below two characteristics.

1. If V1 increases then V2 also increases

2. If V1 decreases then V2 behavior is unknown

A) Pearson correlation will be close to 1  
B) Pearson correlation will be close to -1  
C) Pearson correlation will be close to 0  
D) None of these

**Solution: (D)**

We cannot comment on the correlation coefficient by using only statement 1.  We need to consider the both of these two statements. Consider V1 as x and V2 as |x|. The correlation coefficient would not be close to 1 in such a case.

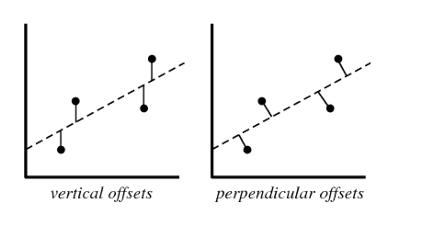
**10) Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?**

A) TRUE  
B) FALSE

**Solution: (B)**

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don’t move together. We can take examples like y=|x| or y=x^2.

**11) Which of the following offsets, do we use in linear regression’s least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.**



A) Vertical offset  
B) Perpendicular offset  
C) Both, depending on the situation  
D) None of above

**Solution: (A)**

We always consider residuals as vertical offsets. We calculate the direct differences between actual value and the Y labels. Perpendicular offset are useful in case of PCA.

**12) True- False: Overfitting is more likely when you have huge amount of data to train?**

A) TRUE  
B) FALSE

**Solution: (B)**

With a small training dataset, it’s easier to find a hypothesis to fit the training data exactly i.e. overfitting.

**13) We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about Normal Equation?**

1. We don’t have to choose the learning rate
2. It becomes slow when number of features is very large
3. Thers is no need to iterate

A) 1 and 2  
B) 1 and 3  
C) 2 and 3  
D) 1,2 and 3

**Solution: (D)**

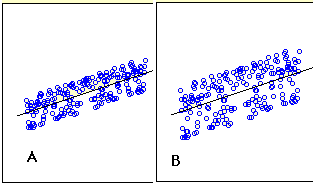
Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this [article](http://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/) for read more about normal equation.

**14) Which of the following statement is true about sum of residuals of A and B?**

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

**Note:**

1. Scale is same in both graphs for both axis.
2. X axis is independent variable and Y-axis is dependent variable.



A) A has higher sum of residuals than B  
B) A has lower sum of residual than B  
C) Both have same sum of residuals  
D) None of these

**Solution: (C)**

Sum of residuals will always be zero, therefore both have same sum of residuals

**Question Context 15-17:**

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penality x.

**15) Choose the option which describes bias in best manner.**  
A) In case of very large x; bias is low  
B) In case of very large x; bias is high  
C) We can’t say about bias  
D) None of these

**Solution: (B)**

If the penalty is very large it means model is less complex, therefore the bias would be high.

**16) What will happen when you apply very large penalty?**

A) Some of the coefficient will become absolute zero  
B) Some of the coefficient will approach zero but not absolute zero  
C) Both A and B depending on the situation  
D) None of these

**Solution: (B)**

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

**17) What will happen when you apply very large penalty in case of Lasso?**  
A) Some of the coefficient will become zero  
B) Some of the coefficient will be approaching to zero but not absolute zero  
C) Both A and B depending on the situation  
D) None of these

**Solution: (A)**

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

**18) 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

**Solution: (A)**

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

**19) Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?**

A) Since the there is a relationship means our model is not good  
B) Since the there is a relationship means our model is good  
C) Can’t say  
D) None of these

**Solution: (A)**

There should not be any relationship between predicted values and residuals. If there exists any relationship between them,it means that the model has not perfectly captured the information in the data.

**Question Context 20-22:**

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is “0” or in another terms it perfectly fits the data.

**20) What will happen when you fit degree 4 polynomial in linear regression?**  
A) There are high chances that degree 4 polynomial will over fit the data  
B) There are high chances that degree 4 polynomial will under fit the data  
C) Can’t say  
D) None of these

**Solution: (A)**

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

**21) What will happen when you fit degree 2 polynomial in linear regression?**  
A) It is high chances that degree 2 polynomial will over fit the data  
B) It is high chances that degree 2 polynomial will under fit the data  
C) Can’t say  
D) None of these

**Solution: (B)**

If a degree 3 polynomial fits the data perfectly, it’s highly likely that a simpler model(degree 2 polynomial) might under fit the data.

**22) In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?**

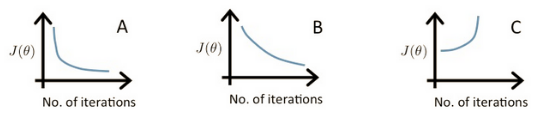
A) Bias will be high, variance will be high  
B) Bias will be low, variance will be high  
C) Bias will be high, variance will be low  
D) Bias will be low, variance will be low

**Solution: (C)**

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

**Question Context 23:**

Which of the following is true about below graphs(A,B, C left to right) between the cost function and Number of iterations?



**23) Suppose l1, l2 and l3 are the three learning rates for A,B,C respectively. Which of the following is true about l1,l2 and l3?**

A) l2 < l1 < l3

B) l1 > l2 > l3  
C) l1 = l2 = l3  
D) None of these

**Solution: (A)**

In case of high learning rate, step will be high, the objective function will decrease quickly initially, but it will not find the global minima and objective function starts increasing after a few iterations.

In case of low learning rate, the step will be small. So the objective function will decrease slowly

**Question Context 24-25:**

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

**24) Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?**

A) Increase  
B) Decrease  
C) Remain constant  
D) Can’t Say

**Solution: (D)**

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

**25) What do you expect will happen with bias and variance as you increase the size of training data?**

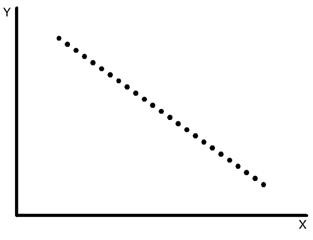
A) Bias increases and Variance increases  
B) Bias decreases and Variance increases  
C) Bias decreases and Variance decreases  
D) Bias increases and Variance decreases  
E) Can’t Say False

**Solution: (D)**

As we increase the size of the training data, the bias would increase while the variance would decrease.

**Question Context 26:**

Consider the following data where one input(X) and one output(Y) is given.



**26) What would be the root mean square training error for this data if you run a Linear Regression model of the form (Y = A0+A1X)?**

A) Less than 0  
B) Greater than zero  
C) Equal to 0  
D) None of these

**Solution: (C)**

We can perfectly fit the line on the following data so mean error will be zero.

**Question Context 27-28:**

Suppose you have been given the following scenario for training and validation error for Linear Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scenario** | **Learning Rate** | **Number of iterations** | **Training Error** | **Validation Error** |
| 1 | 0.1 | 1000 | 100 | 110 |
| 2 | 0.2 | 600 | 90 | 105 |
| 3 | 0.3 | 400 | 110 | 110 |
| 4 | 0.4 | 300 | 120 | 130 |
| 5 | 0.4 | 250 | 130 | 150 |

**27) Which of the following scenario would give you the right hyper parameter?**

A) 1  
B) 2  
C) 3  
D) 4

**Solution: (B)**

Option B would be the better option because it leads to less training as well as validation error.

**28) Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?**

A) Training Error will decrease and Validation error will increase

B) Training Error will increase and Validation error will increase  
C) Training Error will increase and Validation error will decrease  
D) Training Error will decrease and Validation error will decrease  
E) None of the above

**Solution: (D)**

If the added feature is important, the training and validation error would decrease.

**Question Context 29-30:**

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

**29) In such situation which of the following options would you consider?**

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

A) 1 and 2  
B) 2 and 3  
C) 1 and 3  
D) 1, 2 and 3

**Solution: (A)**

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fir the data better.

**30) Now situation is same as written in previous question(under fitting).Which of following regularization algorithm would you prefer?**

A) L1  
B) L2  
C) Any  
D) None of these

**Solution: (D)**

I won’t use any regularization methods because regularization is used in case of overfitting.

**Q1. Which of the following step / assumption in regression modeling impacts the trade-off between under-fitting and over-fitting the most.**

A. The polynomial degree

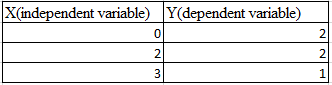
B. Whether we learn the weights by matrix inversion or gradient descent

C. The use of a constant-term

**Solution: A**

Choosing the right degree of polynomial plays a critical role in fit of regression. If we choose higher degree of polynomial, chances of overfit increase significantly.

**Q2. Suppose you have the following data with one real-value input variable & one real-value output variable. What is leave-one out cross validation mean square error in case of linear regression (Y = bX+c)?**



A. 10/27

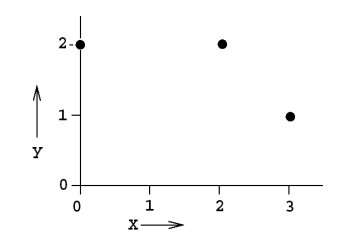
B. 20/27

C. 50/27

D. 49/27

**Solution: D**

We need to calculate the residuals for each cross validation point. After fitting the line with 2 points and leaving 1 point for cross validation.



Leave one out cross validation mean square error = (2^2 +(2/3)^2 +1^2) /3 = 49/27

**Q3. Which of the following is/ are true about  “Maximum Likelihood estimate (MLE)”?**

1. **MLE may not always exist**
2. **MLE always exists**
3. **If MLE exist, it (they) may not be unique**
4. **If MLE exist, it (they) must be unique**

A. 1 and 4

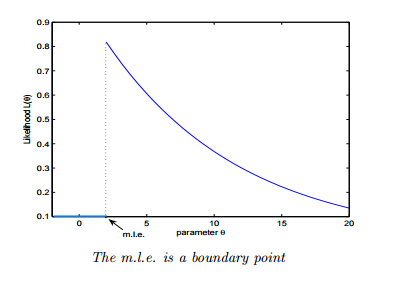
B. 2 and 3

C. 1 and 3

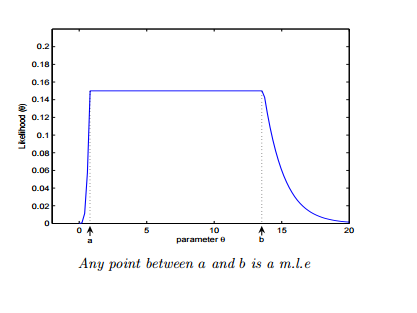
D. 2 and 4

**Solution: C**

The MLE may not be a turning point i.e. may not be a point at which the first derivative of the likelihood (and log-likelihood) function vanishes.



\* The MLE may not be unique.



**Q4. Let’s say, a “Linear regression” model perfectly fits the training data (train error is zero). Now, Which of the following statement is true?**

A. You will always have test error zero

B. You can not have test error zero

C. None of the above

**Solution: C**

Test error may be zero if there no noise in test data. In other words, it will be zero, if the test data is perfect representative of train data but not always.

**Q5. In a linear regression problem, we are using “R-squared” to measure goodness-of-fit. We add a feature in linear regression model and retrain the same model.**

**Which of the following option is true?**

A. If R Squared increases, this variable is significant.

B. If R Squared decreases, this variable is not significant.

C. Individually R squared cannot tell about variable importance. We can’t say anything about it right now.

D. None of these.

**Solution: C**

“R squared” individually can’t tell whether a variable is significant or not because each time when we add a feature, “R squared” can either increase or stay constant. But, it is not true in case of “Adjusted R squared” (increases when features found to be significant).

**Q6. Which one of the statement is true regarding residuals in regression analysis?**

A. Mean of residuals is always zero

B. Mean of residuals is always less than zero

C. Mean of residuals is always greater than zero

D. There is no such rule for residuals.

**Solution: A**

Sum of residual in regression is always zero. It the sum of residuals is zero, the ‘Mean’ will also be zero.

**Q7. Which of the one is true about Heteroskedasticity?**

A. Linear Regression with varying error terms

B. Linear Regression with constant error terms

C. Linear Regression with zero error terms

D. None of these

**Solution: A**

The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises because of presence of outliers or extreme leverage values.

You can refer this [article](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/) for more detail about regression analysis.

**Q8. Which of the following indicates a fairly strong relationship between X and Y?**

A. Correlation coefficient = 0.9

B. The p-value for the null hypothesis Beta coefficient =0 is 0.0001

C. The t-statistic for the null hypothesis Beta coefficient=0 is 30

D. None of these

**Solution: A**

Correlation between variables is 0.9. It signifies that the relationship between variables is fairly strong.

On the other hand, p-value and t-statistics merely measure how strong is the evidence that there is non zero association. Even a weak effect can be extremely significant given enough data.

**Q9. Which of the following assumptions do we make while deriving linear regression parameters?**

1. **The true relationship between dependent y and predictor x is linear**
2. **The model errors are statistically independent**
3. **The errors are normally distributed with a 0 mean and constant standard deviation**
4. **The predictor x is non-stochastic and is measured error-free**

A. 1,2 and 3.

B. 1,3 and 4.

C. 1 and 3.

D. All of above.

**Solution: D**

When deriving regression parameters, we make all the four assumptions mentioned above. If any of the assumptions is violated, the model would be misleading.

**Q10. To test linear relationship of y(dependent) and x(independent) continuous variables, which of the following plot best suited?**

A. Scatter plot

B. Barchart

C. Histograms

D. None of these

**Solution: A**

To test the linear relationship between continuous variables Scatter plot is a good option. We can find out how one variable is changing w.r.t. another variable. A scatter plot displays the relationship between two quantitative variables.

**Q11. Generally, which of the following method(s) is used for predicting continuous dependent variable?**

1. **Linear Regression**
2. **Logistic Regression**

A. 1 and 2

B. only 1

C. only 2

D. None of these.

**Solution: B**

Logistic Regression is used for classification problems. Regression term is misleading here.

**Q12.  A correlation between age and health of a person found to be -1.09.  On the basis of this you would tell the doctors that:**

A. The age is good predictor of health

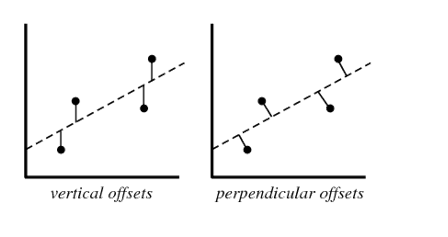
B. The age is poor predictor of health

C. None of these

**Solution: C**

Correlation coefficient range is between [-1 ,1]. So -1.09 is not possible.

**Q13. Which of the following offsets, do we use in case of least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.**



A. Vertical offset

B. Perpendicular offset

C. Both but depend on situation

D. None of above

**Solution: A**

We always consider residual as vertical offsets. Perpendicular offset are useful in case of PCA.

**Q14. Suppose we have generated the data with help of polynomial regression of degree 3 (degree 3 will perfectly fit this data). Now consider below points and choose the option based on these points.**

1. **Simple Linear regression will have high bias and low variance**
2. **Simple Linear regression will have low bias and high variance**
3. **polynomial of degree 3 will have low bias and high variance**
4. **Polynomial of degree 3 will have low bias and Low variance**

A. Only 1

B. 1 and 3

C. 1 and 4

D. 2 and 4

**Solution: C**

If we fit higher degree polynomial greater than 3, it will overfit the data because model will become more complex. If we fit the lower degree polynomial less than 3 which means that we have less complex model so in this case high bias and low variance. But in case of degree 3 polynomial it will have low bias and low variance.

**Q15. Suppose you are training a linear regression model. Now consider these points.**

1. **Overfitting is more likely if we have less data**
2. **Overfitting is more likely when the hypothesis space is small**

Which of the above statement(s) are correct?

A. Both are False

B. 1 is False and 2 is True

C. 1 is True and 2 is False

D. Both are True

**Solution: C**

1.With small training dataset, it’s easier to find a hypothesis to fit the training data exactly i.e. overfitting.

2. We can see this from the bias-variance trade-off. When hypothesis space is small, it has higher bias and lower variance. So with a small hypothesis space, it’s less likely to find a hypothesis to fit the data exactly i.e. underfitting.

**Q16. Suppose we fit “Lasso Regression” to a data set, which has 100 features (X1,X2…X100).  Now, we rescale one of these feature by multiplying with 10 (say that feature is X1),  and then refit Lasso regression with the same regularization parameter.**

**Now, which of the following option will be correct?**

A. It is more likely for X1 to be excluded from the model

B. It is more likely for X1 to be included in the model

C. Can’t say

D. None of these

Solution: B

Big feature values =⇒ smaller coefficients =⇒ less lasso penalty =⇒ more likely to have be kept

**Q17. Which of the following is true about “Ridge” or “Lasso” regression methods in case of feature selection?**

A. Ridge regression uses subset selection of features

B. Lasso regression uses subset selection of features

C. Both use subset selection of features

D. None of above

**Solution: B**

“Ridge regression” will use all predictors in final model whereas “Lasso regression” can be used for feature selection because coefficient values can be zero. For more detail [click here](https://discuss.analyticsvidhya.com/t/difference-between-ridge-regression-and-lasso-and-its-effect/3000).

**Q18. Which of the following statement(s) can be true post adding a variable in a linear regression model?**

1. **R-Squared and Adjusted R-squared both increase**
2. **R-Squared increases and Adjusted R-squared decreases**
3. **R-Squared decreases and Adjusted R-squared decreases**
4. **R-Squared decreases and Adjusted R-squared increases**

A. 1 and 2

B. 1 and 3

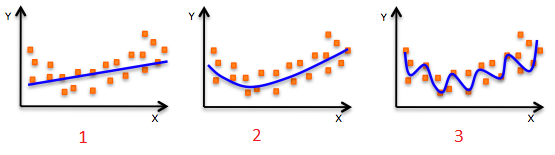
C. 2 and 4

D. None of the above

**Solution: A**

Each time when you add a feature, R squared always either increase or stays constant, but it is not true in case of Adjusted R squared. If it increases, the feature would be significant.

**Q19. The following visualization shows the fit of three different models (in blue line) on same training data. What can you conclude from these visualizations?**



1. **The training error in first model is higher when compared to second and third model.**
2. **The best model for this regression problem is the last (third) model, because it has minimum training error.**
3. **The second model is more robust than first and third because it will perform better on unseen data.**
4. **The third model is overfitting data as compared to first and second model.**
5. **All models will perform same because we have not seen the test data.**

A. 1 and 3

B. 1 and 3

C. 1, 3 and 4

D. Only 5

**Solution: C**

The trend of the data looks like a quadratic trend over independent variable X. A higher degree (Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it under-fits the training data.

**Q20. Which of the following metrics can be used for evaluating regression models?**

1. **R Squared**
2. **Adjusted R Squared**
3. **F Statistics**
4. **RMSE / MSE / MAE**

A. 2 and 4.

B. 1 and 2.

C.  2, 3 and 4.

D. All of the above.

**Solution: D**

These (R Squared, Adjusted R Squared, F Statistics , RMSE / MSE / MAE ) are some metrics which you can use to evaluate your regression model.

**Q21. We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about “Normal Equation”?**

1. **We don’t have to choose the learning rate**
2. **It becomes slow when number of features is very large**
3. **No need to iterate**

A. 1 and 2

B. 1 and 3.

C. 2 and 3.

D. 1,2 and 3.

**Solution: D**

Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this [article](http://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/) for read more about normal equation.

**Q22. The expected value of Y is a linear function of the X(X1,X2….Xn) variables and regression line is defined as:**

**Y = β0 +  β1 X1 + β2 X2……+ βn Xn**

**Which of the following statement(s) are true?**

1. **If Xi changes by an amount ∆Xi, holding other variables constant, then the expected value of Y changes by a proportional amount βi ∆Xi, for some constant βi (which in general could be a positive or negative number).**
2. **The value of βi is always the same, regardless of values of the other X’s.**
3. **The total effect of the X’s on the expected value of Y is the sum of their separate effects.**

**Note: Features are independent of each others(zero interaction).**

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. 1,2 and 3

**Solution: D**

1. The expected value of Y is a linear function of the X variables. This means:
   1. If X i changes by an amount ∆X i , holding other variables fixed, then the expected value of Y changes by a proportional amount β i ∆X i , for some constant β i (which in general could be a positive or negative number).
   2. The value of β i is always the same, regardless of values of the other X’s.
   3. The total effect of the X’s on the expected value of Y is the sum of their separate effects.
2. The unexplained variations of Y are independent random variables (in particular, not “auto correlated” if the variables are time series)
3. They all have the same variance (“homoscedasticity”).
4. They are normally distributed.

**Q23. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?**

A. 1

B. 2

C. Can’t Say

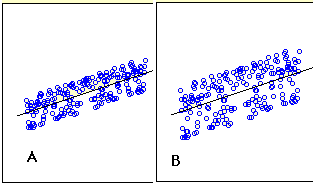
**Solution: B**

In simple linear regression, there is one independent variable so 2 coefficients (Y=a+bx).

**Q24. Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.**

**Note:**

1. **Scale is same in both graphs for both axis.**
2. **X axis is independent variable and Y-axis is dependent variable.**



**Which of the following statement is true about sum of residuals of A and B?**

A) A has higher than B

B) A has lower than B

C) Both have same

D) None of these

**Solution: C**

Sum of residuals always zero.

**Q25. If two variables are correlated, is it necessary that they have a linear relationship?**

A. Yes

B. No

**Solution: B**

It is not necessary. They could have non linear relationship

**Q26. Correlated variables can have zero correlation coeffficient. True or False?**

A. True

B. False

**Solution: A**

**Q27. Suppose I applied a logistic regression model on data and got training accuracy X and testing accuracy Y. Now I want to add few new features in data. Select option(s) which are correct in such case.**

**Note: Consider remaining parameters are same.**

1. **Training accuracy always decreases.**
2. **Training accuracy always increases or remain same.**
3. **Testing accuracy always decreases**
4. **Testing accuracy always increases or remain same**

A. Only 2

B. Only 1

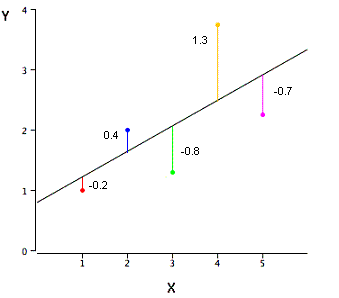
C. Only 3

D. Only 4

Solution: A

Adding more features to model will always increase the training accuracy i.e. low bias. But testing accuracy increases if feature is found to be significant.

**Q28. The graph below represents a regression line predicting Y from X. The values on the graph shows the residuals for each predictions value. Use this information to compute the SSE.**



A. 3.02

B. 0.75

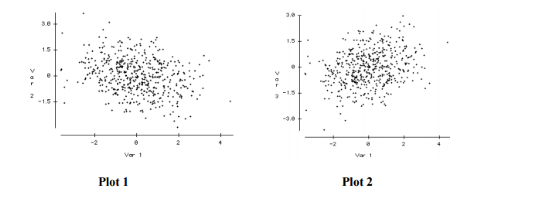
C. 1.01

D. None of these

**Solution: A**

SSE is the sum of the squared errors of prediction, so SSE = (-.2)^2 + (.4)^2 + (-.8)^2 + (1.3)^2 + (-.7)^2 = 3.02

**Q29. Height and weight are well known to be positively correlated. Ignoring the plot scales (the variables have been standardized), which of the two scatter plots (plot1, plot2) is more likely to be a plot showing the values of height (Var1 – X axis) and weight (Var2  – Y axis).**



A. Plot2

B. Plot1

C. Both

D. Can’t say

**Solution: A**

Plot 2 is definitely a better representation of the association between height and weight. As individuals get taller, they take up more volume, which leads to an increase in height, so a positive relationship is expected. The plot on the right has this positive relationship while the plot on the left shows a negative relationship.

**Q30. Suppose the distribution of salaries in a company X has median $35,000, and 25th and 75th percentiles are $21,000 and $53,000 respectively.**

**Would a person with Salary $1 be considered an Outlier?**

A. Yes

B. No

C. More information is required

D. None of these.

**Solution: C**

**Q31. Which of the following option is true regarding “Regression” and “Correlation” ?**

**Note: y is dependent variable and x is independent variable.**

A. The relationship is symmetric between x and y in both.

B. The relationship is not symmetric between x and y in both.

C. The relationship is not symmetric between x and y in case of correlation but in case of regression it is symmetric.

D. The relationship is symmetric between x and y in case of correlation but in case of regression it is not symmetric.

**Solution: D**

1. Correlation is a statistic metric that measures the linear association between two variables. It treats y and x symmetrically.
2. Regression is setup to predict y from x. The relationship is not symmetric.

**Q32. Can we calculate the skewness of variables based on mean and median?**

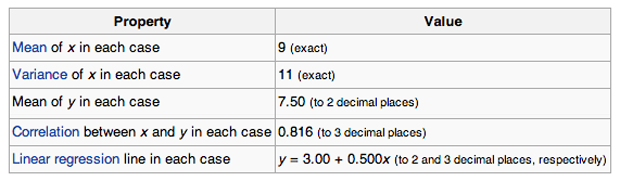
A. True

B. False

**Solution: B**

The skewness is not directly related to the relationship between the mean and median.

**Q33. Suppose you have n datasets with two continuous variables (y is dependent variable and x is independent variable). We have calculated summary statistics on these datasets. All of them give the following result:**



**Are all the given datasets same?**

A. Yes

B. No

C. Can’t Say

**Solutiom: C**

To answer this question, you should know about Anscombe’s quartet. Refer [this link](https://en.wikipedia.org/wiki/Anscombe's_quartet)  to read more about this.

**Q34. How does number of observations influence overfitting? Choose the correct answer(s).**

**Note: Rest all parameters are same**

1. **In case of fewer observations, it is easy to overfit the data.**
2. **In case of fewer observations, it is hard to overfit the data.**
3. **In case of more observations, it is easy to overfit the data.**
4. **In case of more observations, it is hard to overfit the data.**

A. 1 and 4

B. 2 and 3

C. 1 and 3

D. None of theses

**Solution: A**

In particular, if we have very few observations and it’s small, then our models can rapidly overfits data. Because we have only a few points and as we’re increasing in our model complexity like the order of the polynomial, it becomes very easy to hit all of our observations.

On the other hand, if we have lots and lots of observations, even with really, really complex models, it is difficult to overfit because we have dense observations across our input.

**Q35. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option(s) below which describes relationship of bias and variance with lambda.**

A. In case of very large lambda; bias is low, variance is low

B. In case of very large lambda; bias is low, variance is high

C. In case of very large lambda; bias is high, variance is low

D. In case of very large lambda; bias is high, variance is high

**Solution: C**

If lambda is very large it means model is less complex. So in this case bias is high and variance in low.

**Q36. Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with tuning parameter lambda to reduce its complexity. Choose the option(s) below which describes relationship of bias and variance with lambda.**

A. In case of very small lambda; bias is low, variance is low

B. In case of very small lambda; bias is low, variance is high

C. In case of very small lambda; bias is high, variance is low

D. In case of very small lambda; bias is high, variance is high

**Solution: B**

If  lambda is very small it means model is complex. So in this case bias is low and variance is high because model will overfit the data.

**Q37. What is/are true about ridge regression?**

1. **When lambda is 0, model works like linear regression model**
2. **When lambda is 0, model doesn’t work like linear regression model**
3. **When lambda goes to infinity, we get very, very small coefficients approaching 0**
4. **When lambda goes to infinity, we get very, very large coefficients approaching infinity**

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

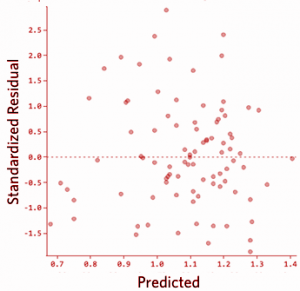
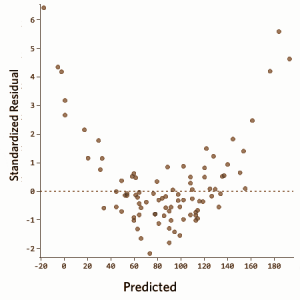
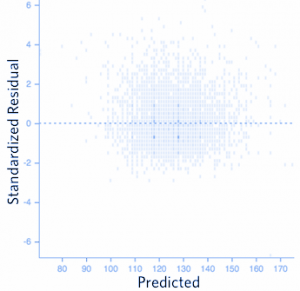
**Solution: A**

Specifically, we can see that when lambda is 0, we get our least square solution. When lambda goes to infinity, we get very, very small coefficients approaching 0.

**Q38. Out of the three residual plots given below, which of the following represent worse model(s) compared to others?**

**Note:**

1. **All residuals are standardized.**
2. **The plots are between predicted values Vs. residuals**

A. 1

B. 2

C. 3

D. 1 and 2

**Solution: C**

There should not be any relationship between predicted values and residuals. If there exist any relationship between them means model has not perfectly capture the information in data.

**Q39. Which of the following method(s) does not have closed form solution for its coefficients?**

A. Ridge regression

B. Lasso

C. Both Ridge and Lasso

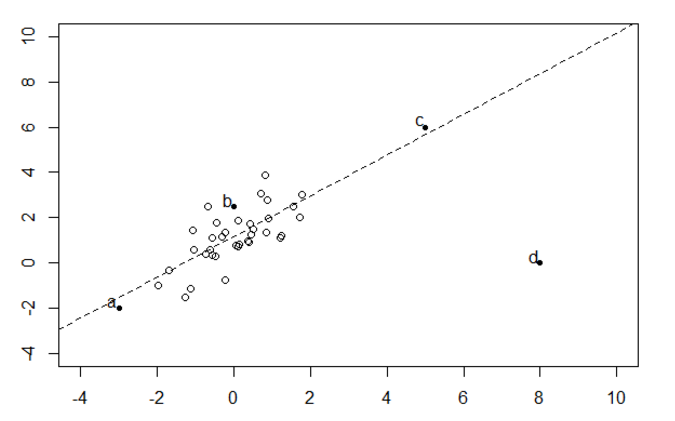
D. None of both

**Solution: B**

The Lasso does not admit a closed-form solution. The L1-penalty makes the solution non-linear. So we need to approximate the solution.

If you want to read more about closed form solutions, refer this [link](http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf).

**Q40.  Consider the following dataset**



**Which bold point, if removed will have the largest effect on fitted regression line as shown in above figure(dashed)?**

A) a

B) b

C) c

D) d

**Solution: D**

Linear regression is sensitive to outliers in the data. Although c is also an outlier in given data space but it is closed to the regression line(residual is less) so it will not affect much.

**Q41. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?**

A: By 1

B. No change

C. By intercept

D. By its Slope

**Solution: D**

Equation for simple linear regression: Y=a+bx. Now if we increase the value of x by 1 then the value of y would be a+b(x+1) i.e. value of y will get incremented by b.

**Q42. Logistic Regression transforms the output probability to be in a range of [0, 1]. Which of the following function is used by logistic regression to convert the probability in the range between [0,1].**

A. Sigmoid

B. Mode

C. Square

D. Probit

**Solution: A**

Sigmoid function is used to convert output probability between [0,1] in logistic regression.

**Q43: Which of the following statement is true about partial derivative of the cost functions w.r.t weights / coefficients in linear-regression and logistic-regression?**

A. Both will be different

B. Both will be same

C. Can’t say

D. None of these

**Solution: B**

Refer this [link](http://feature-space.com/2011/10/28/logistic-cost-function-derivative/)

**Q44. Suppose, we are using Logistic regression model for n-class classification problem. In this case, we can use One-vs-rest method. Choose which of the following option is true regarding this?**

A. We need to fit n model in n-class classification problem.

B. We need to fit n-1 models to classify into n classes.

C. We need to fit only 1 model to classify into n classes.

D. None of these.

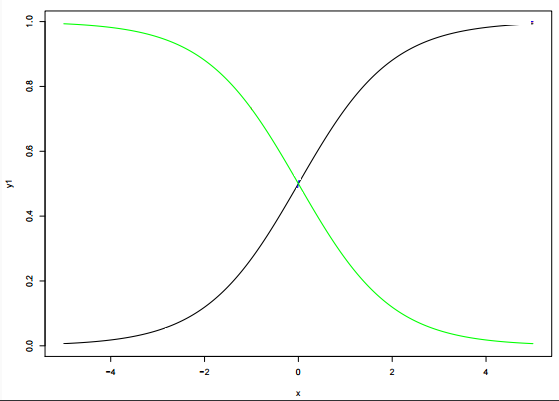
**Solution: A**

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

Take a example of 3-class(-1,0,1) classification. Then need to train 3 logistic regression classifiers.

1. -1 vs 0 and 1
2. 0 vs -1 and 1
3. 1 vs 0 and -1

**Q45. Below are two different logistic models with different values for β0 and β1.**



**Which of the following statement(s) is true about β0 and β1 values of two logistics models (Green, Black)?**

**Note: consider Y = β0 + β1\*X.  Here, β0 is intercept and  β1 is  coefficient.**

A. β1 for Green is greater than Black

B. β1 for Green is lower than Black

C. β1 for both models is same

D. Can’t Say.

**Solution: B**

β0 and β1: β0 = 0, β1 = 1 is in X1 color(black) and β0 = 0, β1 = −1 is in X4 color (green)

**1) Which of the following is/are true about bagging trees?**

1. In bagging trees, individual trees are independent of each other
2. Bagging is the method for improving the performance by aggregating the results of weak learners

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

**2) Which of the following is/are true about boosting trees?**

1. In boosting trees, individual weak learners are independent of each other
2. It is the method for improving the performance by aggregating the results of weak learners

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: B**

In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.

**3) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?**

1. Both methods can be used for classification task
2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
4. Both methods can be used for regression task

A) 1  
B) 2  
C) 3  
D) 4  
E) 1 and 4

**Solution: E**

Both algorithms are design for classification as well as regression task.

**4) In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest?**

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

**Solution: A**

Random forest is based on bagging concept, that consider faction of sample and faction of feature for building the individual trees.

**5) Which of the following is true about “max\_depth” hyperparameter in Gradient Boosting?**

1. Lower is better parameter in case of same validation accuracy
2. Higher is better parameter in case of same validation accuracy
3. Increase the value of max\_depth may overfit the data
4. Increase the value of max\_depth may underfit the data

A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

**Solution: A**

Increase the depth from the certain value of depth may overfit the data and for 2 depth values validation accuracies are same we always prefer the small depth in final model building.

**6) Which of the following algorithm doesn’t uses learning Rate as of one of its hyperparameter?**

1. Gradient Boosting
2. Extra Trees
3. AdaBoost
4. Random Forest

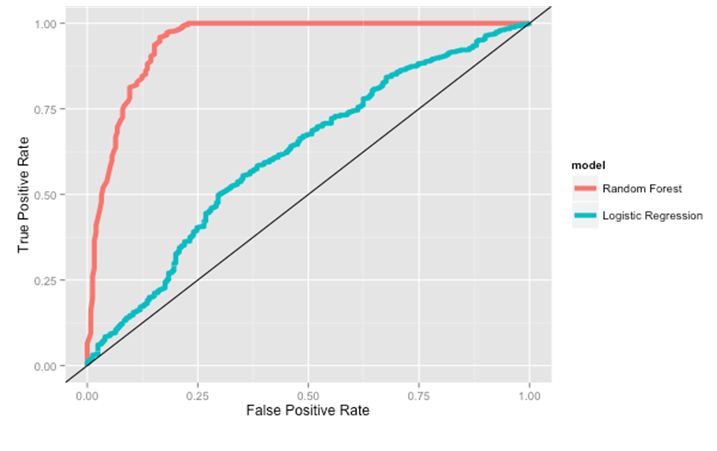
A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

**Solution: D**

Random Forest and Extra Trees don’t have learning rate as a hyperparameter.

**7) Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?**

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/07/12181501/7_image_dt.jpg)A) Random Forest  
B) Logistic Regression  
C) Both of the above  
D) None of these

**Solution: A**

Since, Random forest has largest AUC given in the picture so I would prefer Random Forest

**8) Which of the following is true about training and testing error in such case?**

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training T1, T2 … Tn where T1 < T2…. Tn-1 < Tn.

A) The difference between training error and test error increases as number of observations increases  
B) The difference between training error and test error decreases as number of observations increases  
C) The difference between training error and test error will not change  
D) None of These

**Solution: B**

As we have more and more data, training error increases and testing error de-creases. And they all converge to the true error.

**9) In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?**

A) Only Random forest algorithm handles real valued attributes by discretizing them  
B) Only Gradient boosting algorithm handles real valued attributes by discretizing them  
C) Both algorithms can handle real valued attributes by discretizing them  
D) None of these

**Solution: C**

Both can handle real valued features.

**10) Which of the following algorithm are not an example of ensemble learning algorithm?**

A) Random Forest  
B) Adaboost  
C) Extra Trees  
D) Gradient Boosting  
E) Decision Trees

**Solution: E**

Decision trees doesn’t aggregate the results of multiple trees so it is not an ensemble algorithm.

**11) Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?**

1. Number of tree should be as large as possible
2. You will have interpretability after using RandomForest

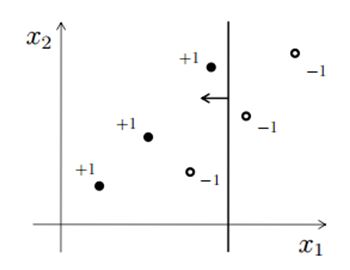
A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: A**

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building.  Random Forest is a black box model you will lose interpretability after using it.

**Context 12-15**

Consider the following figure for answering the next few questions. In the figure, X1 and X2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X1(say splitting point is x11) which is shown in the figure using vertical line. Every value less than x11 will be predicted as positive class and greater than x will be predicted as negative class.

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/07/14112922/12_image_con.jpg)**12) How many data points are misclassified in above image?**

A) 1  
B) 2  
C) 3  
D) 4

**Solution: A**

Only one observation is misclassified, one negative class is showing at the left side of vertical line which will be predicting as a positive class.

**13) Which of the following splitting point on feature x1 will classify the data correctly?**

A) Greater than x11  
B) Less than x11  
C) Equal to x11  
D) None of above

**Solution: D**

If you search any point on X1 you won’t find any point that gives 100% accuracy.

**14) If you consider only feature X2 for splitting. Can you now perfectly separate the positive class from negative class for any one split on X2?**

A) Yes  
B) No

**Solution: B**

It is also not possible.

**15) Now consider only one splitting on both (one on X1 and one on X2) feature. You can split both features at any point. Would you be able to classify all data points correctly?**

A) TRUE  
B) FALSE

**Solution: B**

You won’t find such case because you can get minimum 1 misclassification.

**Context 16-17**

Suppose, you are working on a binary classification problem with 3 input features. And you chose to apply a bagging algorithm(X) on this data. You chose max\_features = 2 and the n\_estimators =3. Now, Think that each estimators have 70% accuracy.

Note: Algorithm X is aggregating the results of individual estimators based on maximum voting

**16) What will be the maximum accuracy you can get?**

A) 70%  
B) 80%  
C) 90%  
D) 100%

**Solution: D**

Refer below table for models M1, M2 and M3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Actual predictions** | **M1** | **M2** | **M3** | **Output** |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 | 1 |

**17) What will be the minimum accuracy you can get?**

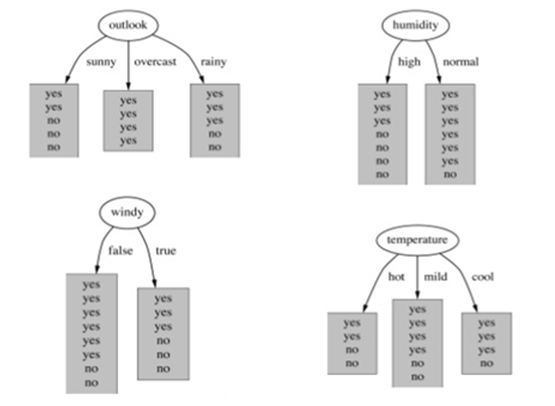
A) Always greater than 70%  
B) Always greater than and equal to 70%  
C) It can be less than 70%  
D) None of these

**Solution: C**

Refer below table for models M1, M2 and M3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Actual predictions** | **M1** | **M2** | **M3** | **Output** |
| 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 |

**18) Suppose you are building random forest model, which split a node on the attribute, that has highest information gain. In the below image, select the attribute which has the highest information gain?**

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/07/14115929/19_Image.jpg)  
A) Outlook  
B) Humidity  
C) Windy  
D) Temperature

**Solution: A**

Information gain increases with the average purity of subsets. So option A would be the right answer.

**19) Which of the following is true about the Gradient Boosting trees?**

1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model
2. We can use gradient decent method for minimize the loss function

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

Both are true and self explanatory

**20) True-False: The bagging is suitable for high variance low bias models?**

A) TRUE  
B) FALSE

**Solution: A**

The bagging is suitable for high variance low bias models or you can say for complex models.

**21) Which of the following is true when you choose fraction of observations for building the base learners in tree based algorithm?**

A) Decrease the fraction of samples to build a base learners will result in decrease in variance  
B) Decrease the fraction of samples to build a base learners will result in increase in variance  
C) Increase the fraction of samples to build a base learners will result in decrease in variance  
D) Increase the fraction of samples to build a base learners will result in Increase in variance

**Solution: A**

Answer is self explanatory

**Context 22-23**

Suppose, you are building a Gradient Boosting model on data, which has millions of observations and 1000’s of features. Before building the model you want to consider the difference parameter setting for time measurement.

**22) Consider the hyperparameter “number of trees” and arrange the options in terms of time taken by each hyperparameter for building the Gradient Boosting model?**

Note: remaining hyperparameters are same

1. Number of trees = 100
2. Number of trees = 500
3. Number of trees = 1000

A) 1~2~3  
B) 1<2<3

C) 1>2>3  
D) None of these

**Solution: B**

The time taken by building 1000 trees is maximum and time taken by building the 100 trees is minimum which is given in solution B

**23) Now, Consider the learning rate hyperparameter and arrange the options in terms of time taken by each hyperparameter for building the Gradient boosting model?**

Note: Remaining hyperparameters are same

1. learning rate = 1  
2. learning rate = 2  
3. learning rate = 3

A) 1~2~3  
B) 1<2<3

C) 1>2>3  
D) None of these

**Solution: A**

Since learning rate doesn’t affect time so all learning rates would take equal time.

**24) In greadient boosting it is important use learning rate to get optimum output. Which of the following is true abut choosing the learning rate?**

A) Learning rate should be as high as possible  
B) Learning Rate should be as low as possible  
C) Learning Rate should be low but it should not be very low  
D) Learning rate should be high but it should not be very high

**Solution: C**

Learning rate should be low but it should not be very low otherwise algorithm will take so long to finish the training because you need to increase the number trees.

**25) [True or False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.**

A) TRUE  
B) FALSE

**Solution: A**

**26) When you use the boosting algorithm you always consider the weak learners. Which of the following is the main reason for having weak learners?**

1. To prevent overfitting
2. To prevent under fitting

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: A**

To prevent overfitting, since the complexity of the overall learner increases at each step. Starting with weak learners implies the final classifier will be less likely to overfit.

**27) To apply bagging to regression trees which of the following is/are true in such case?**

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bias

A) 1 and 2  
B) 2 and 3  
C) 1 and 3  
D) 1,2 and 3

**Solution: D**

All of the options are correct and self explanatory

**28) How to select best hyperparameters in tree based models?**

A) Measure performance over training data  
B) Measure performance over validation data  
C) Both of these  
D) None of these

**Solution: B**

We always consider the validation results to compare with the test result.

**29) In which of the following scenario a gain ratio is preferred over Information Gain?**

A) When a categorical variable has very large number of category  
B) When a categorical variable has very small number of category  
C) Number of categories is the not the reason  
D) None of these

**Solution: A**

When high cardinality problems, gain ratio is preferred over Information Gain technique.

**30) Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario** | **Depth** | **Training Error** | **Validation Error** |
| 1 | 2 | 100 | 110 |
| 2 | 4 | 90 | 105 |
| 3 | 6 | 50 | 100 |
| 4 | 8 | 45 | 105 |
| 5 | 10 | 30 | 150 |

A) 1  
B) 2  
C) 3  
D) 4

**Solution: B**

Scenario 2 and 4 has same validation accuracies but we would select 2 because depth is lower is better hyper parameter.