Quiz 1

1. Using too large a value of lambda() in regularization can cause your hypothesis to underfit the data.
   Options:
   a. **True**
   b. False

2. Both being tree-based algorithms, how is random forest different from Gradient boosting algorithm (GBM)?
   Ans:
   The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.

3. What is the difference between boosting and bagging?
   Ans:
   Bagging, bootstrap aggregation: using the data itself and using all of them. Bagging creates multiple copies of the original training dataset using the bootstrap, fitting a model to each copy, and then combining all of the models in order to create a single predictive model.
   Boosting: Boosting tries to add new models that do well where previous models fail.

4. Which of the following algorithm are not an example of ensemble learning algorithm?
   Options:
   a. Random forest
   b. GBM
   c. AdaBoost
   d. Decision Tree
   e. XGBoost

5. Which of the following is/are true about ensemble models?
   Options:
   a. In boosting trees, individual weak learners are independent of each other.
   b. Random Forest doesn't have learning rate as of one of its hyperparameter
   c. Ensembles models give better interpretability
   d. The bagging is suitable for high variance low bias models or you can say for complex models.

6. What is ROC? Is ROC suitable for regression or classification? What is the relationship between ROC and AUC?
   Ans:
   The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. ROC is suitable for classification. AUC is the area under the ROC curve

7. Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. First, we use model A and get an error rate of 20% on the training data and 25% on the test data. Then, we use model B and get an error rate of 22.5% on the training data and 22.5% on the test data.

   Which model is preferred? Why?

   Ans:
   We care about the error rate on the test data, so we would prefer model B.

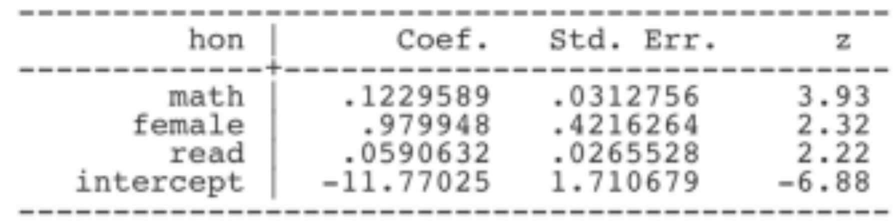
8. How do you achieve Bias-variance tradeoff for data which does not fit well?
   Options:
   a. Regularization
   b. Bagging
   c. Boosting
   d. **All of the above**

9. Bias is error on __________ data, while variance is error on ___________ data.
   Options:
   a. **Training, testing**
   b. Testing, training
   c. Training, training
   d. Testing, testing

10. What are the the important tuning parameters of supervised machine learning algorithm Support Vector Machines (SVM)
    Ans:
    Kernel: helps in finding separation lines in higher dimensions. Linear kernel, polynomial kernel
    Regularization: tuning parameter (C) used to control misclassification of samples. Large C, smaller margin separation, small C, larger margin separation.

Gamma: low gamma far away points help in finding separation line, high gamma, nearby points help in finding separation line.

11. Assume regression is being used to predict whether a student will graduate with honors or not. The dependent variable is hon. A yes as indicated by a 1 a no indicated by a 0. Assume that we have three independent variables: 1.) one called math and is an integer representing a student's math score. 2.) one called read and is an integer representing a student's reading score. And 3) a gender variable called female where the value (female = 1) means the gender is female. The stats for the fit are shown in the table below.



```
---------------------------------------------------
          hon |      Coef.    Std. Err.        z
--------------+------------------------------------
         math |   .1229589    .0312756      3.93
       female |    .979948    .4216264      2.32
         read |   .0590632    .0265528      2.22
    intercept |  -11.77025    1.710679     -6.88
---------------------------------------------------
```

Write an equation that describes the model.

For all of the fitted parameters answer the following:

- Is the coefficient significant?
- How does one interpret the meaning of its value?

Ans:

This fitted model says that, holding math and reading at a fixed value, the odds of getting into an honors class for females (female = 1) over the odds of getting into an honors class for males (female = 0) is exp(.979948) = 2.66. In terms of percent change, we can say that the odds for females are 166% higher than the odds for males. The coefficient for math says that, holding female and reading at a fixed value, we will see 13% increase in the odds of getting into an honors class for a one-unit increase in math score since exp(.1229589) = 1.13.

12. Given a bag having 5 red and 3 black balls:

i. What are the odds in favor of having a red ball in the bag?

ii. Similarly, what are the odds not in favor of having a red ball (rather an odd of having a black ball)?

iii. What is the probability of having a red ball?

iv. Is odds and probability one and the same?

Options:

a. 3/5, 5/3, 5/8, False

b. 5/8, 3/8, 5/8, True

c. **5/3, 3/5, 5/8, False**
d. None of the above

13. What is a standardized regression coefficient? How do we calculate them?

    Ans:

    They are coefficients on the same scale.
    Subtract mean and divide by the standard deviation.

14. What problem a large k value will cause in k-fold cross validation?

    Ans:

    Too large k means that only a low number of sample combinations is possible, thus limiting the number of iterations that are different, which means more variance and lower bias.

15. Describe an interaction variable in regression.

    Ans:

    An interaction effect exists when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables.

16. Interaction effect is represented as:

    Options:

    a. **Product of two or more independent variables.**
    b. Product of two or more dependent variables
    c. Sum of the independent variables
    d. Constant term

17. How does one know whether the effect of interaction variables is significant?

    Options:

    a. P-value
    b. Z-score
    c. T-score
    d. **All the above**

18. Feature selection:

    Options:

    a. enables the machine learning algorithm to train faster
    b. Increases the complexity of a model
    c. improves the accuracy of a model
    d. All of the above

19. Which of the following are incorrect about dummy variables?

Options:

a. Dummy variables assign the numbers ‘0’ and ‘1’ to indicate membership in any mutually exclusive and exhaustive category.

b. The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable.

c. The interaction of two attribute variables (e.g. Gender and Marital Status) is represented by a third dummy variable which is simply the product of the two individual dummy variables.

20. Write the following regression equation with interaction

$$\hat{y} = b0 + b1X1 + b2X2$$

Ans:

$\hat{y} = b0 + b1X1 + b2X2 + b3X1X2$