Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8401: APPLIED MACHINE LEARNING IN DATA SCIENCE

DATE: 17th October 2022 Time: 2 Hours

Instructions

1. Total Points: 100
2. Answer **ALL** questions in Part A, Part B and Part C.
3. Answer any **ONE** question in Part D.

Part A: True or False Section (Provide a brief justification for your choice)

1. (2 pts) The individual trees in a random forest are all trained on all of the training data. True or False? Explain your choice:

2. (2 pts.) Principal Components Analysis (PCA) and Sequential Forward Selection are two examples of unsupervised methods for dimensionality reduction. True or False? Explain your choice:

3. (2 pts) Minimizing the entropy is equivalent to maximizing the information gain. True or False? Explain your choice:

4. (2 pts) Regularization in linear regression produces larger coefficients and hence steeper model slopes. True or False? Explain your choice:

5. (2 pts) When a decision tree is grown to full depth, it is likely to be more accurate and also less likely to fit the noise in the data. True or False? Explain your choice:

6. (2 pts) Normalizing the data is an effective strategy to reduce overfitting. True or False? Explain your choice:

7. (2 pts) Gradient descent is guaranteed to find the global minimum. True or False? Explain your choice:

8. (2 pts) Pruning in decision trees is a method to prevent overfitting. True or False? Explain your choice:

9. (2 pts) Grid search hyper-parameter optimization process is an efficient method for finding the value of the coefficients in Logistic Regression. True or False? Explain your choice:

10. (2 pts) In Gradient Boosting, misclassified individual data points are up-weighted from one tree to the next. True or False? Explain your choice:

Part B: Multiple Choice Section

Part B1: For the following questions, check (✓) ALL CORRECT CHOICES. The set of all correct answers must be checked. **No partial credit**!

1. (3 pts) What strategies can help reduce overfitting in decision trees?
   a. Pruning
   b. Make sure each leaf node is one pure class
   c. Enforce a minimum number of samples in leaf nodes
   d. Enforce a maximum depth for the tree
2. (3 pts) Which of the following are true about sequential forward feature selection?
   a. It greedily adds the feature that most improves accuracy
   b. It finds the subset of features that give the lowest test error
   c. Forward selection is faster than backward selection if few features are relevant to prediction
   d. Is an iterative method wherein we start with a full model i.e., all features in the model
3. (3 pts) Which of the following statements are correct?
   a. Machine Learning is good for solutions requiring long list of rules
   b. Machine Learning is good for fluctuating environments e.g., data changes, problem changes
   c. Machine Learning is good when the cost of error is too high and every decision or action needs to be explainable
   d. Machine Learning is good for dealing with large, complex data
4. (3 pts) Which of the following statements about Precision and Recall are correct?
   a. If Precision is high and Recall is high, then F1 Score is HIGH
   b. If Precision is low and Recall is low, then F1 Score is LOW
   c. If one is high and the other is low, F1 Score is MEDIUM
   d. Precision and Recall have no impact on F1 Score
5. (3 pts) Assume 2 decision trees are trained on the same data and same algorithm settings. Which of the following statements are correct?

   a. A tree with depth of 3 has higher variance than a tree with depth of 1.
   b. A tree with depth of 3 has higher bias than a tree with depth 1.
   c. A tree with depth of 3 never has higher training error than a tree with depth 1.
   d. A tree with depth of 3 never has higher test error than a tree with depth 1.

Part B2: For the following questions, check (✓) THE CORRECT CHOICE. Only **ONE** choice is correct

1. (3 pts) Which of the following constitutes a Machine Learning task
   a. Supervised Learning
   b. Unsupervised Learning
   c. Both the above
   d. None of the above
2. (3 pts)..................... is a widely used and effective machine learning algorithm based on the idea of bagging
   a. Regression
   b. Classification
   c. Decision Tree
   d. Random Forest
3. (3pts) What is one disadvantage of decision trees?
   a. Cannot handle non-linearity
   b. Decision trees are robust to outliers
   c. Decision trees are prone to be overfit
   d. All of the above
4. (3 pts) How can you handle missing data in a dataset?
   a. Drop missing rows or columns
   b. Assign a unique category to missing values
   c. Replace missing values with mean/median/mode
   d. All of the above
5. (3 pts) Which of the following are metrics and tools to assess a classification model?
   a. Confusion matrix
   b. Precision and Recall
   c. Area under the ROC curve
   d. All of the above
6. (3 pts) A Machine Learning technique that helps in detecting the outliers in data.
   a. Association Rule Mining
   b. Classification
   c. Isolation Forest
   d. All of the above
7. (3 pts) Which of the following is not a supervised learning?
   a. PCA
   b. Ridge Regression
   c. Linear Regression
   d. Decision Tree
8. (3 pts) Data used to optimize the hyperparameter settings of a supervised model is called ....
   a. Test
   b. Training
   c. Validation
   d. None of the above
   e. any of the above

9. (3 pts) When choosing one feature from X1, ..., Xn while building a Decision Tree, which of the following criteria is the most appropriate to maximize? (Here, H() means an entropy, and P() means a Probability)
    a. P(Y |Xj)
    b. P(Y) − P(Y |Xj)
    c. H(Y) − H(Y |Xj)
    d. H(Y |Xj)
10. (3 pts) You've just finished training a random forest for spam classification, and it is getting abnormally bad performance on your validation set, but good performance on your training set. Your implementation has no bugs. What could be causing the problem?
    a. Your decision trees are too deep
    b. You are randomly sampling too many features when you choose a split
    c. You have too few trees in your ensemble
    d. Your bagging implementation is randomly sampling sample points without replacement

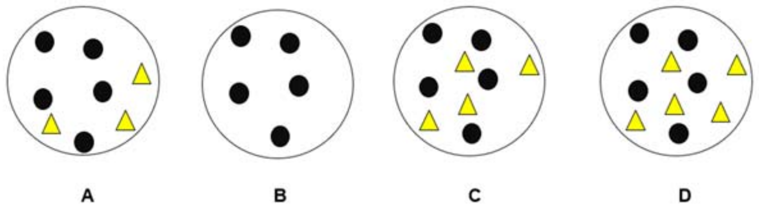
Part C: Short Answer Section

1. (2 pts.) The ROC Curve plots the ……………………………………rate against the …………………………rate.

2. (2 pts.) …………………………………… and ………………………… are two examples of packages/approaches that can be used for model explainability/interpretation.

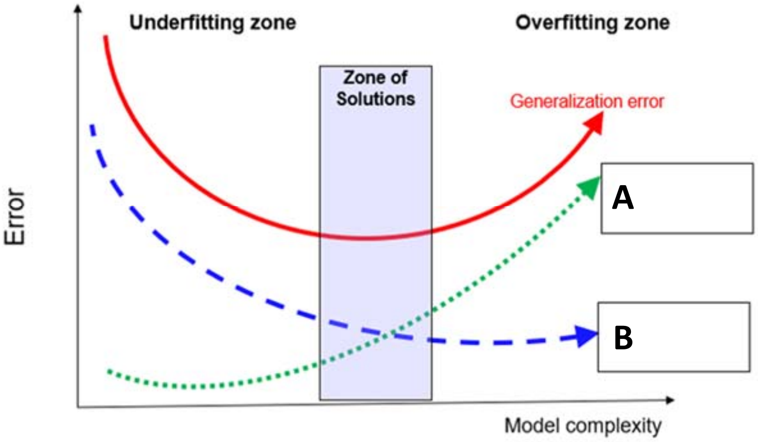
3. (2 pts) Which one of the three main classes of feature selection methods does NOT require information about the target feature?

4. (2 pts) Given the 5 datasets below which are a mix of circles and triangles. Arrange the datasets in order of decreasing Entropy, from highest to lowest.



5. (2 pts) Name 2 regularization techniques that can be applied to Logistic Regression

6. (2 pts) There are three theoretical error sources in supervised modeling. Variance Error, Bias Error and Irreducible Error. In the chart below, please label the blank boxes with the types of error represented by the 2 unlabeled lines:
    a. A =
    b. B =



7. (2 pts) Under-performing classifiers can fall into one of 2 regimes:
    a. Name 1 remedy for a High Variance regime
    b. Name 1 remedy for a High Bias regime

8. (3 pts) Give one example of a hyperparameter and one example of a model parameter encountered in machine learning.

9. (3pts) List one difference between Adaptive Boosting (Adaboost) and Gradient Boosting.

10. (3 pts) Name 3 reasons why model explainability is important.

11. (3 pts) Name 3 hyperparameters that can be tuned while building Random Forest Models.

12. (3 pts) When building an SVM model, name a situation where you would need to apply the Kernel trick.

13. (2 pts) Consider the accuracy vs explainability tradeoff. Name 1 model type that can be considered 'black box' and 1 model type that can be considered inherently explainable ('glass box').

Part D: Choice Section (Answer any 1 question)

1. (4 pts) A student obtained a modeling dataset where one of the features was ***'Gender'*** which was coded as **F** female and **M** for male. The student wanted to convert this feature to numerical so she created a new feature called '***Gender_new'*** where she assigned a value of **1** for females and **2** for males. Ok or Problematic? If Problematic, briefly state what the problems are:

2. (4 pts) A student claimed great success after achieving 99% classification accuracy on a fraud classification task where their data consisted of 100 positive examples and 9,900 negative examples. Ok or Problematic? If Problematic, briefly state what the problems are:

3. (4 pts) You work for the Food Standards Board. You are building a binary classifier that detects whether food sold in a store contains salmonella. If your system doesn't catch infected food, somebody will get sick. Should you optimize for the highest/lowest accuracy, precision, or recall? Why?