Fast**National University of Computer & Emerging Sciences, Karachi  
Fall-2023 FAST School of Computing  
Final Exam**

**2nd January 2024, 09:00 am – 12:00 noon**

**SOLUTION**

|  |  |  |
| --- | --- | --- |
| **Course Code:** CS4101 | **Course Name:** Applied Machine Learning | |
| **Instructor Name:** Dr. M. Shahzad | | |
| **Student Roll No:** | | **Section:** |

**Instructions:**

* Return the question paper and keep it inside your answer sheet.
* Read each question completely before answering it. There are **three questions and two pages (front plus back)**.
* In case of any ambiguity, you may make assumptions. However, your assumption should not contradict any statement in the question paper.
* Do not write anything on the question paper (except your ID and group).

**Total Time:** 1 Hour **Max Points**: 15

------------------------------------------------------------------------------------------------------------------------

**Question 1: [XAI] [5 points, CLO2, 20 mins]**

1. The managing board of a corporation consists of three stockholders who have respectively 80, 60, and 30 shares of stock and the chairman who has no shares. Any decision can be settled by approval of board members holding a simple majority of the shares and the chairman can decide tie votes. We number the players as follows: stockholder with 80 shares—Player 1, a stockholder with 60 shares—Player 2, a stockholder with 30 shares—Player 3. Thus, we define a characteristic function of the coalition as follows:

v({1,2}) = 180, v({1,3}) = 160, v({2,3}) = 120, v({1,2,3}) = 260.

1. Find the Shapley value of each player. Show all calculations.

**Solution:**

The Shapley value is Player 1 = 115, Player 2 = 85, and Player 3 = 60 and is calculated as follows:

A table with numbers and text

Description automatically generated

1. Is Player 1 a dummy player?

**Solution:**

Player 1 is not a dummy player, because v({1,2}) - v({2}) = 180 – 60 = 120 > v({1}) = 80.

1. Are Players 1 and 2 interchangeable?

**Solution:**

Players 1 and 2 are not interchangeable because v({1}) ≠ v({2}).

1. **Give short answers to the questions. An answer exceeding 03 (THREE) lines will not be marked.**
2. Suppose you are working on a dataset with three features: X1, X2 and X1 + X2. How would you determine the SHAP values for just singular features? Would it be correct in saying that the true SHAP value of X1 is SHAP(X1 + X2) - SHAP(X2) due to SHAP values being linear?

**Solution:**

No. SHAP values are calculated based on the marginal contribution of a given feature to the prediction, considering all possible combinations of features. Therefore, if you have combined features (such as X1 + X2) in your dataset, their contributions can't necessarily be easily separated, so you can't really state that SHAP(X1) = SHAP(X1 + X2) - SHAP(X2). Stating that SHAP(X1) = SHAP(X1 + X2) - SHAP(X2) wouldn't be mathematically sound as it oversimplifies the interdependencies between features.

1. The **marginal contribution** is defined either as “*the difference between the prediction with and without the feature”* or *“the difference between the feature effect minus the average effect”.* Differentiate both statements.

**Solution:**

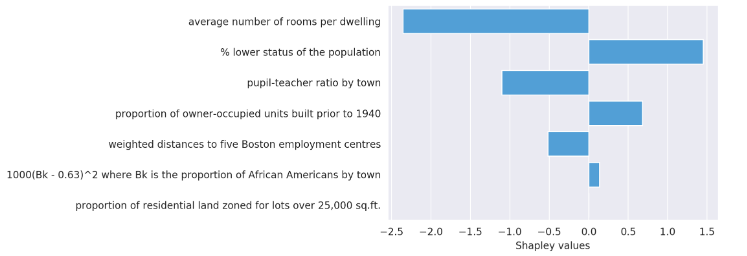
The first definition reflects the importance of a feature for a single prediction (i.e., local), while the second reflects the (entire) model's performance (i.e., global).

1. How can the Shapley value be applied in the context of resource allocation? Give a scenario, application of shapely value, and benefits in the same context.

**Solution:**

* **Scenario:** Consider a group of individuals collaborating on a project, and each individual contributes different skills or resources to the overall success of the project.
* **Application of Shapley Value:** The Shapley value can be used to fairly distribute the benefits or rewards of the project among the contributors. It provides a principled way to allocate the total value generated by the collaboration to each individual based on their marginal contributions to different possible coalitions.
* **Benefits:** This ensures that each contributor receives a share of the benefits that reflect their unique contribution to the collective effort.

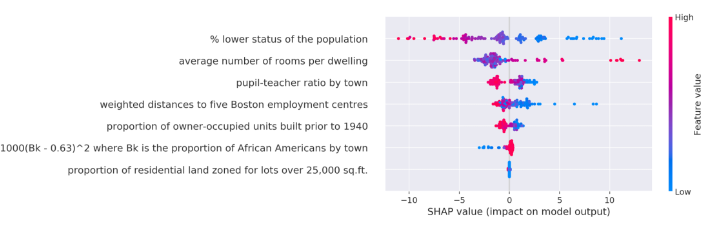
1. Suppose you apply SHAP on Boston House Prices data set. *Figure 1* shows the Shapley value of each feature for the randomly picked instance, ranked by their absolute importance. Give briefly intuitive sense of the two most important feature.



**Solution:**

* + The average number of rooms negatively affects the price prediction i.e. these houses offer fewer rooms and hence are expected to be cheaper.
  + The second most important feature is the population which has a lower social status. This feature is difficult to read in the way it is constructed, however, the model learns that a high percentage of African Americans in a neighborhood lowers the average house price. This can lead to racial bias.

1. Consider Figure 3 below which shows a summary plot of the same Boston House Prices data:



* 1. Which feature is most important, and how does this graph differ from the one you observed in part 4 of this question?

**Solution:** the status of the population is more important for the prediction than the average number of rooms per dwelling. Because it gives a global interpretation.

* 1. Identify the least significant feature and how you interpret it.

**Solution:** the residential land zoned for lots over 25,000 sq.ft.is the most unimportant feature because all the instances’ shap values are 0’s.

**Question 2: [MLOps] [5 points, CLO2, 20 mins]**

**Give short answers to the following questions. An answer exceeding 03 (THREE) lines will not be marked.**

1. What is the role of the GitHub Actions and Docker Hub in CI/CD pipeline?

**Solution:** GitHub Actions automates workflows including build and test processes, while Docker Hub stores Docker images, both playing vital roles in CI/CD pipelines for continuous integration and delivery.

1. Differentiate between data drift and concept drift.

**Solution:** **Data drift:** change in input data and **Concept drift:** change in the relationship between input and output data

1. In your MLOps workshop, how were containers utilized within the context of CI/CD pipelines?

**Solution:** They used containers in (CI/CD) pipeline to provide a consistent and portable environment for developing, testing, and deploying applications of their project on the Portuguese banking Dataset.

1. In the context of MLOps, what functionalities does Streamlit tool offer?

**Solution:** Streamlit enables the creation of user-friendly interfaces for MLOps projects, facilitating real-time predictions and enhancing user interaction with ML models.

1. What did you hear about Kubernetes in the MLOps session?

**Solution:** Kubernetes is an open-source container orchestration platform that automates the deployment, scaling, and management of containerized applications.

**Question 3: [AutoML] [5 points, CLO2, 20 mins]**

**Give short answers to the following questions. An answer exceeding 03 (THREE) lines will not be marked.**

1. Can we use AutoML for a small dataset e.g. 100 data points only? If yes, then how?

**Solution:** Yes, but with some limitations like only using LDA or a limited part of it.

1. What prompted **Prof. Frank Hutter** to address the hyperband and successive halving algorithms in the context of AutoML?

**Solution:** Approximate blackboxes or optimizing hyperparameters and model configurations

1. Do you agree that Hyperband is better than successive halving when resources are limited? Provide the rationale for your preference.

**Solution:** Yes. The bracketing mechanism in Hyperband enables the algorithm to allocate resources optimally. Instead of uniformly allocating resources across configurations, Hyperband focuses resources on configurations that show promise, making it more resource-efficient.

1. How a random grid search combined with the stacked ensemble is a powerful combination?

**Solution:** The combination of a random grid search and a stacked ensemble leverages the benefits of efficient hyperparameter exploration and the strength of combining diverse models, leading to a powerful and adaptable approach for model building and optimization.

1. How H2o AutoML leaderboard work?

**Solution:** H2o AutoML produces a leaderboard that ranks the trained model based on a predefined metric. By default, it ranks models by ascending order of logloss and rmse for classification and regression tasks respectively.

**Question 4: [Time Series Analysis] [5 points, CLO2, 20 mins]**

**Give short answers to the following questions.**

1. Give an example of the statement: "Forecasting is a kind of prediction, but not all predictions are forecasts".

**Solution:**  **Weather prediction** is forecasting as the prediction is specific, data-driven, and involves anticipating future weather patterns based on systematic analysis. Whereas **Sales prediction** is not forecasting but prediction because it doesn't qualify as a forecast in the same rigorous sense as weather forecasting.

1. What are the main components of an AR model? How do you identify it?

**Solution:** An AR model is represented with set of coefficients that represent the weights assigned to the past observations. In an AR(p) model, where "p" is the order of the autoregressive process, these coefficients (ϕ1, ϕ2, ..., ϕp) determine the relationship between the current observation and its "p" lagged values. The lag order, denoted as "p," signifies the number of past observations included in the autoregressive model. It determines the extent to which the current value of the time series is influenced by its past values. Selecting an appropriate lag order is crucial for model accuracy. The Autocorrelation Function (ACF) measures the correlation between a time series and its lagged values. In the context of identifying an AR model, the ACF plot helps identify the order of the autoregressive process. If there is a significant correlation at lag "k" and subsequent lags, it suggests an AR(k) model will best represents the given time series data.

**Question 5: [Association Rule Mining] [5 points, CLO2, 20 mins]**

Trace the results of using the Apriori algorithm on the grocery store example with support threshold **s=33.34%** and confidence threshold **c=60%**. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

|  |  |
| --- | --- |
| Transaction ID | Items |
| T1 | Kabab, Buns, Ketchup |
| T2 | Kabab, Buns |
| T3 | Kabab, Coke, Chips |
| T4 | Chips, Coke |
| T5 | Chips, Ketchup |
| T6 | Kabab, Coke, Chips |

Solution:

Support threshold =33.34% => threshold is at least 2 transactions. Applying Apriori

|  |  |  |
| --- | --- | --- |
| Pass (k) | Candidate k-itemsets and their support | Frequent k-itemsets |
| K=1 | Kabab(4), Buns(2), Ketchup(2), Coke(3), Chips(4) | Kabab, Buns, Ketchup, Coke, Chips |
| K=2 | {Kabab, Buns}(2), {Kabab, Ketchup}(1), {Kabab, Coke}(2), {Kabab, Chips}(2), {Buns, Ketchup}(1), {Buns, Coke}(0), {Buns, Chips}(0), {Ketchup, Coke}(0), {Ketchup, Chips}(1), {Coke, Chips}(3) | {Kabab, Buns}, {Kabab, Coke}, {Kabab, Chips}, {Coke, Chips} |
| K=3 | {Kabab, Coke, Chips}(2) | {Kabab, Coke, Chips} |
| K=4 | {} |  |

*Note that {Kabab, Buns, Coke} and {Kabab, Buns, Chips} are not candidates when k=3 because their subsets {Buns, Coke} and {Buns, Chips} are not frequent.*

*Note also that normally, there is no need to go to k=4 since the longest transaction has only 3 items.*

**All Frequent Itemsets:**

{Kabab}, {Buns}, {Ketchup}, {Coke}, {Chips}, {Kabab, Buns}, {Kabab,

Coke}, {Kabab, Chips}, {Coke, Chips}, {Kabab, Coke, Chips}.

**Association rules:**

{Kabab, Buns} would generate: Kabab -> Buns (2/6=0.33, 2/4=0.5) and

Buns -> Kabab (2/6=0.33, 2/2=1);

{Kabab, Coke} would generate: Kabab -> Coke (0.33, 0.5) and

Coke -> Kabab (2/6=0.33, 2/3=0.66);

{Kabab, Chips} would generate: Kabab -> Chips (0.33, 0.5) and

Chips -> Kabab (2/6=0.33, 2/4=0.5);

{Coke, Chips} would generate: Coke -> Chips (3/6=0.5, 3/3=1) and

Chips -> Coke (3/6=0.5, 3/4=0.75);

{Kabab, Coke, Chips} would generate: Kabab -> Coke ^ Chips (2/6=0.33, 2/4=0.5),

Coke -> Chips ^ Kabab (2/6=0.33, 2/3=0.66),

Chips -> Coke ^ Kabab (2/6=0.33, 2/4=0.5),

Kabab ^ Coke -> Chips(2/6=0.33, 2/2=1),

Kabab ^ Chips -> Coke(2/6=0.33, 2/2=1) and

Coke ^ Chips -> Kabab(2/6=0.33, 2/3=0.66).

**With the confidence threshold set to 60%, the Strong Association Rules are (sorted by confidence):**

1. Coke -> Chips (0.5, 1)

2. Buns -> Kabab (0.33, 1);

3. Kabab ^ Coke -> Chips(0.33, 1)

4. Kabab ^ Chips -> Coke(0.33, 1)

5. Chips -> Coke (0.5, 0.75);

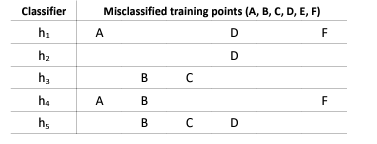
6. Coke -> Kabab (0.33, 0.66);

7. Coke -> Chips ^ Kabab (0.33, 0.66)

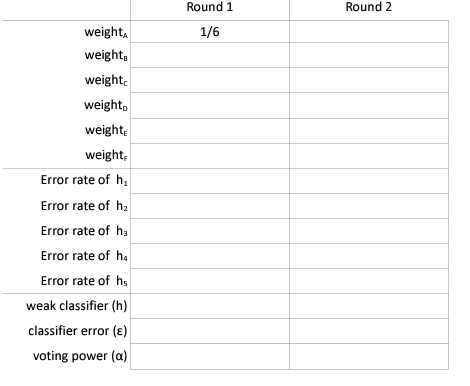
8. Coke ^ Chips -> Kabab (0.33, 0.66).

**Question 6: [Ensemble Learning] [5 points, CLO2, 20 mins]**

You have six training points (A, B, C, D, E, F) and five classifiers (h1, h2, h3, h4, h5) which make the following misclassifications:



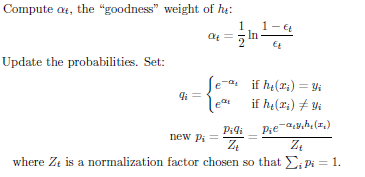
Perform two rounds of boosting with these classifiers and training data. In each round, pick the classifier with the lowest error rate. Break ties by picking the classifier that comes first in this list: h1, h2, h3, h4, h5. Give your results of each round in the following format:



Formulas you may require:

A close up of a sign

Description automatically generated



**Question 7: [Neural Network] [5 points, CLO2, 20 mins]**

1. Let's assume we have a simple function *f(x, y, z)* = *(x+y)z*. We can break this up into the equations *q = x+y* and *f(x, y, z) = qz*. Using this simplified notation, we can also represent this equation as a computation graph:

A diagram of a diagram

Description automatically generated

Now let's assume that we are evaluating this function at **x = -2​, y = 5​, and z = -4​.** In addition, let the value of the upstream gradient (gradient of the loss with respect to our function, ) equal 1. These are filled out for you in the computation graph.

**Task**: Solve for the following values, both symbolically (without plugging in specific values of x/y/z), and evaluated **at x = -2​, y = 5​, z = -4​,** and ∂L/∂f = 1​:

A table of mathematical equations

Description automatically generated

**Solution:**

A diagram of a mathematical equation

Description automatically generated with medium confidence

1. Complete the following table by applying Perceptron for 2 iterations. Start with random weights of w1 = 1, w2 = -1. Clearly mention the new weights. Pseudocode of updated rule is also given. For iteration 2, pick the first example that is misclassified.

A white background with black text

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input1 | Input2 | Actual Output | Predicted Output (Iteration 1) | Predicted Output (Iteration 2) |
| 2 | 3 | 1 |  |  |
| 0 | 4 | -1 |  |  |
| 1 | -1 | 1 |  |  |

**Solution:**

Apply current weights.

Iteration # 1: 2x1 + 3x-1 = 2-3 = -1; 0x1 + 4x-1 = -4 <0 = -1; 1x1 + 1x-1 = 0 >=0 = 1

Now pick 1st example and update weights

w1 = w1 + 2 = 1+2 = 3; w2 = -1+3 = 2

Iteration # 2: 2x3 + 3x2 = 6+6 = 12; 0x3 + 4x2 = 8 > 0 = 1; 1x3 + 4x2 = 8 >=0 = 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input1 | Input2 | Actual Output | Predicted Output (Iteration 1) | Predicted Output (Iteration 2) |
| 2 | 3 | 1 | -1 | 1 |
| 0 | 4 | -1 | -1 | 1 |
| 1 | -1 | 1 | 1 | 1 |

**Question 8: [Naïve Bayes] [5 points, CLO2, 20 mins**

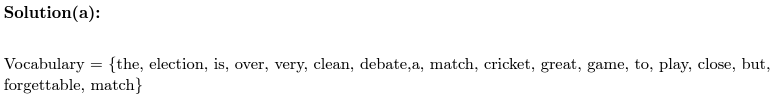
This problem is related to the Naive Bayes classifier for text classification that helps us in classifying the given statement as ‘Political’ or ‘Not political’.

Consider the following training and test data. We will build a Naive Bayes classifier using a bag-of-words approach to assign labels to the last two test statements.

A table with words on it

Description automatically generated

1. **(2pts)** Formulate a vocabulary based on the provided training data.



1. **(4pts)** Construct a bag-of-words representation for each class utilizing the training data.

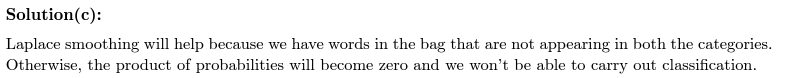
**Hint: Create a table as per the following format:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Bag of Words** | **Total** | **Political** | **Non-Political** |
| The | 2 | 2 | 0 |
| Election | 2 | 2 | 0 |





1. **(3pts)** Is Laplace Smoothing necessary for both the given training and test data? Present a concise justification to support your response.

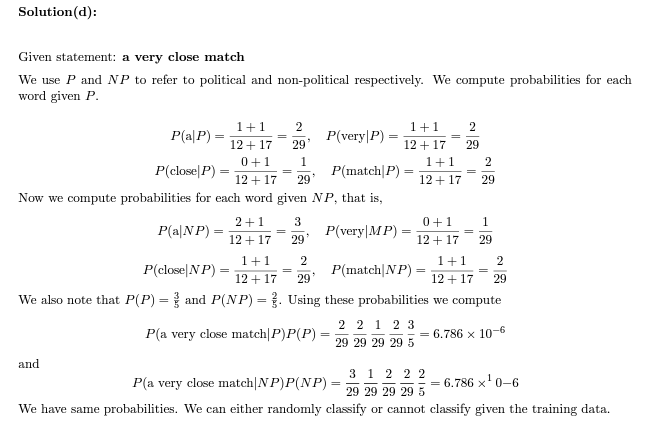


1. **(6pts)** Employ the Naive Bayes approach to classify the given test statements as 'Political' or 'Not Political.' This involves computing P(Political).P(statement | Political) and P(Not Political).P(statement | Not Political) for each test statement. Utilize add-1 smoothing for the calculation of probabilities. Laplace smoothing formula is given below:

A close up of a number

Description automatically generated

**Hint:** You need to split the statement into words to estimate *P(statement|Class) e.g. P(I am Paksitani|True) = P(I|True).P(am|True).P(Pakistani|True)*



A math equations and numbers

Description automatically generated with medium confidence

**BEST OF LUCK!**