Quiz 4 - Supervised Learning Algorithms

* Indicates required question

1. Name *

2. Email *

Tree based Models

3. What is the primary criterion used to find the optimal split in a Classification Decision Tree? 1 point

Algorithm Decision Tree Learning Algorithm

Require: Training data $\{(F_i, y_i)\}_{i=1}^{n}$, stopping criteria
Ensure: Decision tree T

Initialize tree with single root node containing all data
while nodes can be split and stopping criteria not met **do**
  for each leaf node with region $\mathcal{R}$ **do**
    Find (j^*, τ^*) that maximizes:
    $IG(j, \tau) = I(\mathcal{R}) - \frac{|\mathcal{R}_L|}{|\mathcal{R}|} I(\mathcal{R}_L) - \frac{|\mathcal{R}_R|}{|\mathcal{R}|} I(\mathcal{R}_R)$
    Where $\mathcal{R}_L = \{F \in \mathcal{R} : F_j \leq \tau\}$ and $\mathcal{R}_R = \{F \in \mathcal{R} : F_j > \tau\}$
    Split node using rule $F_{j^*} > \tau^*$
  end for
end while
Assign prediction to each leaf node (majority class)
return T

Mark only one oval.

- Minimize Mean Squared Error
- Maximize Information Gain
- Minimize Sum of Squared Errors
- Gradient Descent

4. In a Regression Tree, what is the optimization objective when choosing a feature and threshold for splitting? 1 point

Algorithm Regression Tree Learning Algorithm

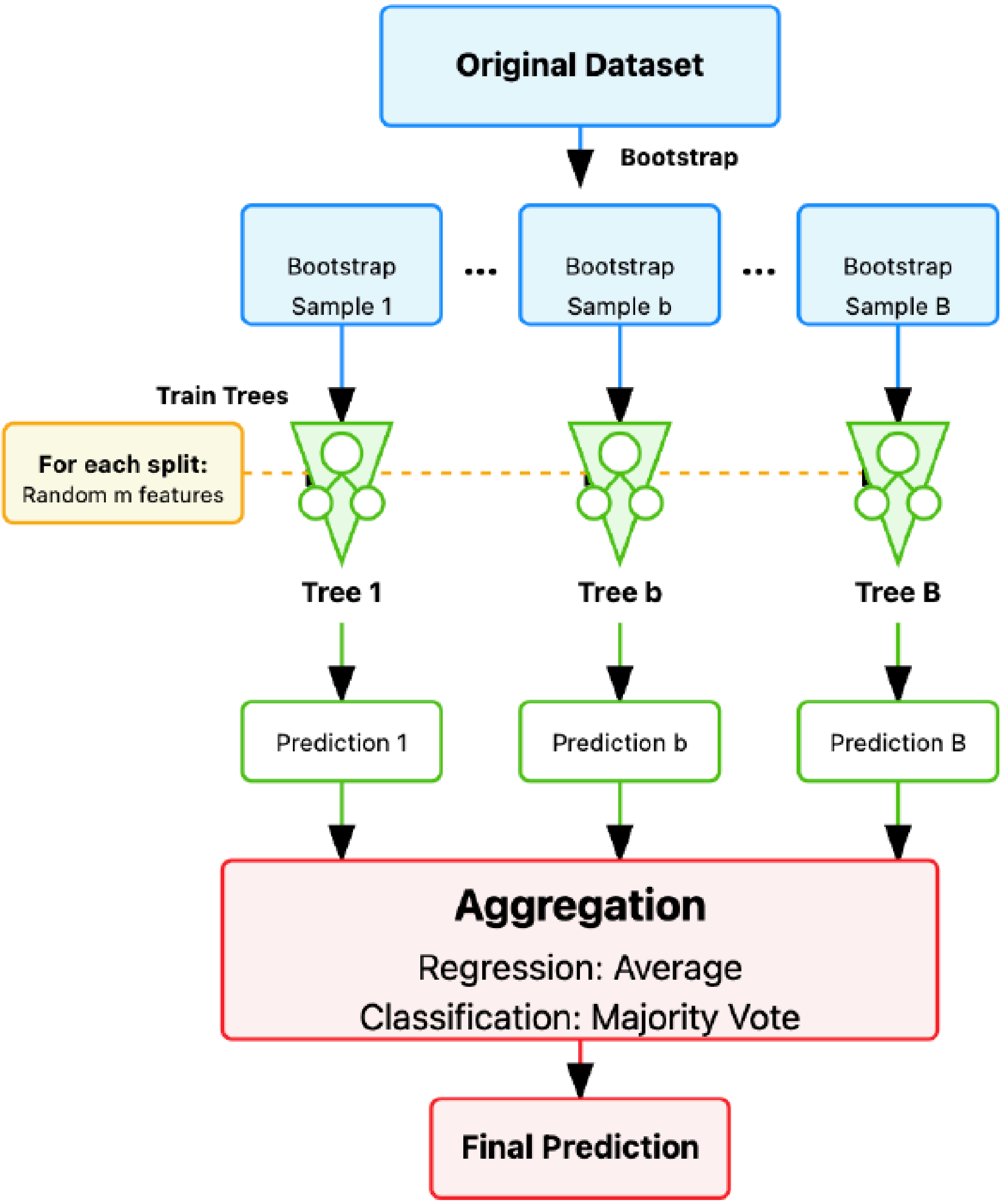
Require: Training data $\{(F_i, y_i)\}_{i=1}^{n}$, stopping criteria
Ensure: Regression tree T

```
Initialize tree with single root node containing all data
while nodes can be split and stopping criteria not met do
    for each leaf node with region R do
        Find (j*, τ*) that minimizes:
        SSE(j, τ) = Σ_{i:F_i∈R_L} (y_i − ȳ_{R_L})² + Σ_{i:F_i∈R_R} (y_i − ȳ_{R_R})²
        Where R_L = {F ∈ R : F_j ≤ τ} and R_R = {F ∈ R : F_j > τ}
        Split node using rule F_{j*} > τ*
    end for
end while
Assign prediction ȳ_{R_m} to each leaf node (average of y_i in the region)
return T
```

Mark only one oval.

- Maximize Information Gain
- Minimize Sum of Squared Errors in resulting regions
- Maximize the number of samples in each region
- Minimize the entropy in each region

5. How does Random Forest differ from standard bagging of decision trees? 1 point



Mark only one oval.

- ⬭ Random Forest uses subsets of features at each split
- ⬭ Random Forest only uses a single decision tree with random initialization
- ⬭ Random Forest only applies to regression problems
- ⬭ Random Forest eliminates bootstrapping entirely

6. **True/False**: In AdaBoost, if a weak learner has an error rate of exactly 0.5 (random guessing), its contribution weight (α) will be zero. 1 point

Mark only one oval.

- True
- False

7. For the binary classification with labels {0, 1}, the weight update formula in AdaBoost can be written as: 1 point

Mark only one oval.

$$w_{t+1}(i) \propto w_t(i) \cdot e^{\alpha_t y_i f_t(x_i)}$$

A

$$w_{t+1}(i) \propto w_t(i) \cdot e^{-\alpha_t y_i f_t(x_i)}$$

B

$$w_{t+1}(i) \propto w_t(i) \cdot e^{\alpha_t(1-2\cdot\mathbb{I}\{y_i=f_t(x_i)\})}$$

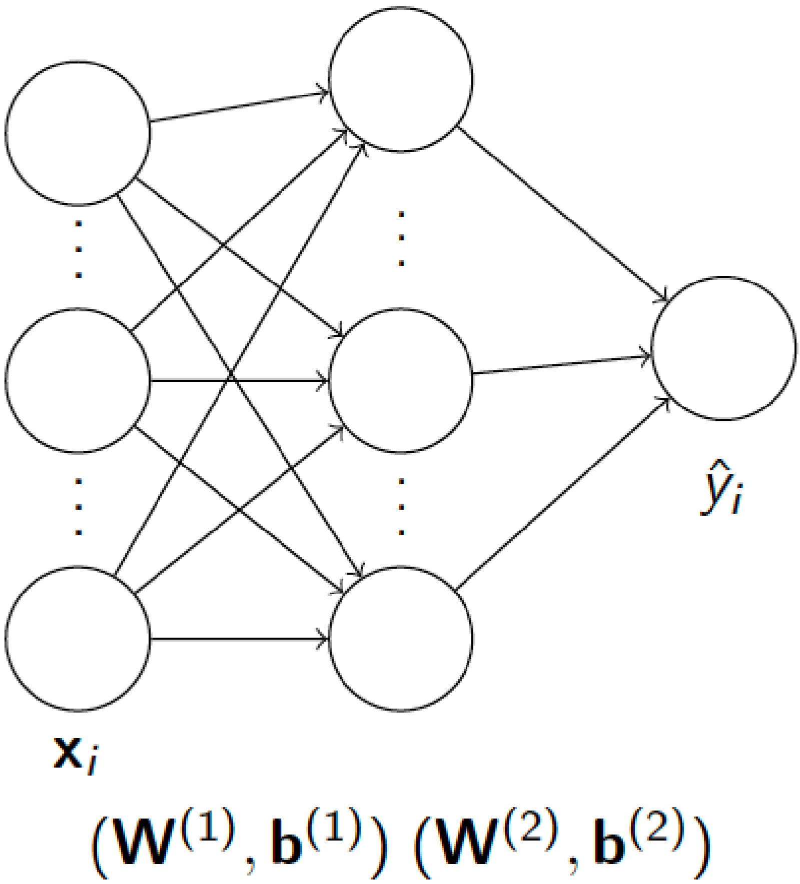
C

$$w_{t+1}(i) \propto w_t(i) \cdot e^{-\alpha_t(1-2\cdot\mathbb{I}\{y_i=f_t(x_i)\})}$$

D

Neural Networks

8. In a shallow neural network with one hidden layer, what are the components of the parameter set θ? 1 point



Mark only one oval.

- Only the weights connecting input to hidden layer
- Only the weights connecting hidden to output layer
- Weights connecting input to hidden layer and hidden to output layer
- Weights and biases for both hidden and output layers

9. Which activation function has the range [0,1] and is commonly used as the output activation for binary classification problems? 1 point

Mark only one oval.

- ReLU
- Tanh
- Sigmoid
- ELU

10. For a multi-class classification problem with K classes, which loss function and final layer activation would typically be used? 1 point

Mark only one oval.

- Mean Squared Error with linear activation
- Binary Cross-Entropy with sigmoid activation
- Categorical Cross-Entropy with softmax activation
- Mean Absolute Error with tanh activation

11. In a regression task with potential outliers in the data, which loss function would be most appropriate? 1 point

Mark only one oval.

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE) or Huber Loss
- Binary Cross-Entropy
- Categorical Cross-Entropy

12. In the Gradient Descent algorithm, what happens if the learning rate is set too high? 1 point

Algorithm Gradient Descent Algorithm

Require: Training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, loss function $\mathcal{L}$, learning rate α, iterations T

Ensure: Optimized parameters θ

```
Initialize parameters θ^(0) randomly
for t = 1 to T do
    θ^(t) = θ^(t−1) − α · ∇_θ L(θ^(t−1))
    if Convergence criteria met then
        break
    end if
end for
return Final parameters θ^(T)
```

Mark only one oval.

- ◯ Convergence will be very slow
- ◯ The algorithm may overshoot the minimum and potentially diverge
- ◯ The algorithm will always converge to the global minimum

Questions

13. Any comment ?

____