BA476

Yunzhe Yu

February 2024

Problem 1

(a) The logistic regression model is given by:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X1 + \beta_2 X2$$

where $\beta_0 = 0.6$, $\beta_1 = 3$, and $\beta_2 = -0.1$.

Solving for p (the probability) gives us:

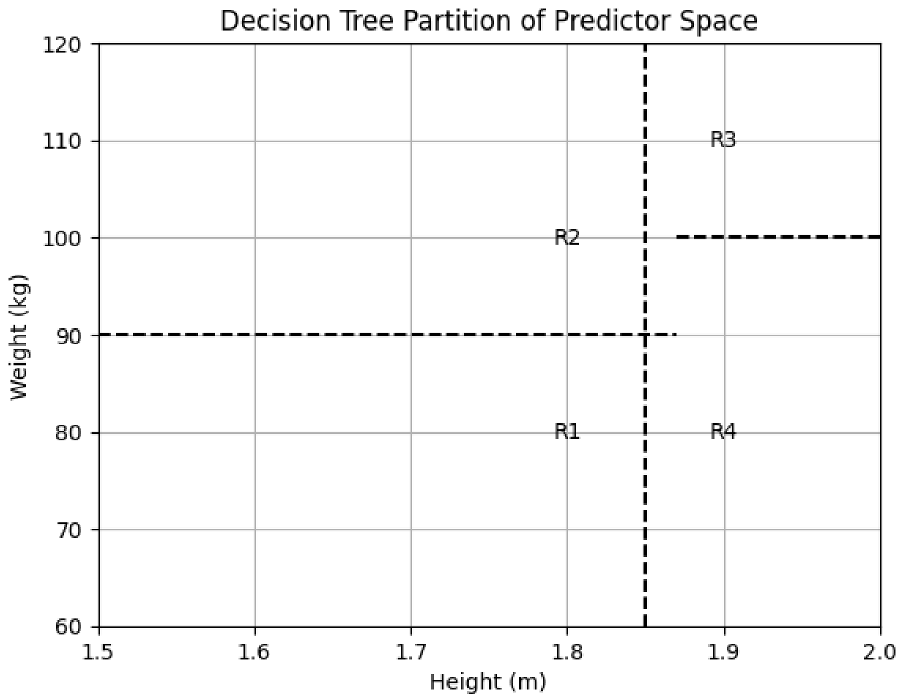
$$p = \frac{e^{\beta_0+\beta_1 X1+\beta_2 X2}}{1+e^{\beta_0+\beta_1 X1+\beta_2 X2}}$$

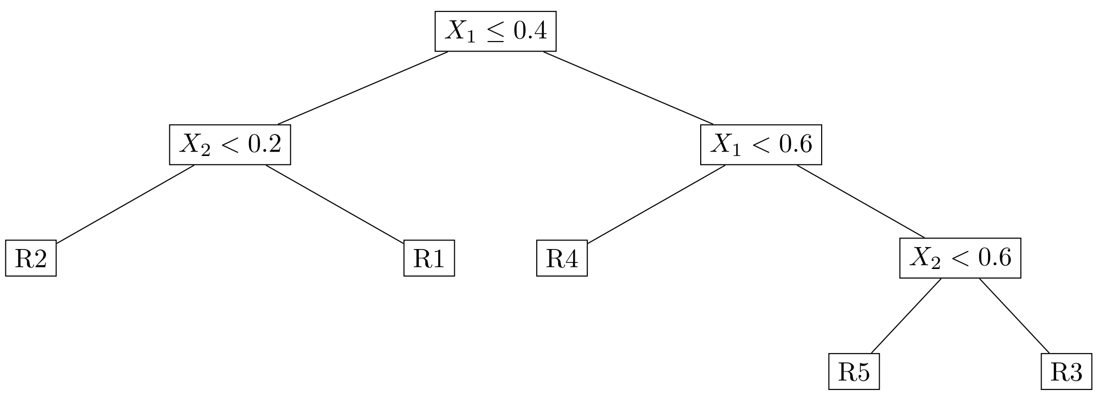
Using a threshold $t = 0.5$, we find the accuracy of the model to be 80%.

(b) To find an alternative threshold to maximize accuracy, we look for a threshold between the incorrectly predicted probability (approximately 0.52) and the closest higher correct probability (approximately 0.79). An alternative threshold at the midpoint of these values is approximately 0.655.

(c) For a predicted probability of $p(x) = 0.7$, the odds of observing label 1 are given by $\frac{p(x)}{1-p(x)}$, which is approximately 2.33.

(d) For the final part, since $e^{\beta_1} = e^3$ increases the odds, the statement that having $X1 = 1$ increases your odds of having label 1 by $\beta_1 = 3$ is "True."


Decision Tree Partition of Predictor Space
Weight (kg)
Height (m)
R1
R2
R3
R4
120
110
100
90
80
70
60
1.5
1.6
1.7
1.8
1.9
2.0


$X_1 \leq 0.4$
$X_2 < 0.2$
$X_1 < 0.6$
R2
R1
R4
$X_2 < 0.6$
R5
R3

(a)

Problem 3

- Tree (a) has 3 misclassification.
- Tree (b) has 4 misclassifications.
- Tree (c) has 2 misclassifications.

(a) The initial split of Tree (c) makes the fewest mistakes as it has the lowest number of misclassifications (2).

(b) Number of instances incorrectly classified in each of the three trees:

- Tree (a) misclassifies 3 instances.
- Tree (b) misclassifies 4 instances.
- Tree (c) misclassifies 2 instances.

(c) The number of splits $|T|$ in each tree:

- Tree (a) has 2 splits.
- Tree (b) has 1 split.
- Tree (c) has 0 splits.

(d) The tree selected by cost-complexity pruning when $\alpha = 0.5$ is Tree (c).

(e) The tree selected by cost-complexity pruning when $\alpha = 2$ is also Tree (c).

Problem 4

- For the first plot (leftmost), where the classes are separated by a curve, we can use a parabolic feature:

  $$X_3 = X_1^2 + X_2^2$$

  This will allow a decision stump to split the classes using a threshold on X_3.

- For the second plot (middle), with classes linearly separable by a diagonal line, we define:

  $$X_3 = X_1 - X_2$$

  A decision stump can then classify the instances by applying a threshold on X_3.

- For the third plot (rightmost), where the classes are arranged in a circular pattern:

  $$X_3 = \sqrt{X_1^2 + X_2^2}$$

  Here, X_3 represents the Euclidean distance from the center of the circle, allowing for a radial separation.

Problem 5

1. The first tree f_1 is trained on a bootstrap sample selected by the sequence (5, 1, 6, 5, 2, 3) and utilizes 'Age' and 'Tightness_in_chest' as predictors. The decision tree finds the optimal split at an 'Age' of 56 years. The rules are as follows:

   - If Age ≤ 56: predict `False` for Cardiac Arrest.
   - If Age > 56: predict `True` for Cardiac Arrest.

2. The second tree f_2 is trained on a bootstrap sample selected by the sequence (3, 2, 1, 2, 6, 5) and uses 'ECG' and 'Tightness_in_chest' as predictors. The decision tree finds the optimal split using the 'ECG' predictor. The rules are:

   - If ECG ≤ 0.5: predict `False` for Cardiac Arrest (corresponds to 'Normal').
   - If ECG > 0.5: predict `True` for Cardiac Arrest (corresponds to 'Hypertrophy' or 'Abnormal').

3. The third tree f_3 is trained on a bootstrap sample selected by the sequence (2, 2, 4, 1, 1, 5) and uses 'Bp_change' and 'Tightness_in_chest' as predictors. Despite the presence of a split on 'Bp_change', the tree concludes that all instances predict a cardiac arrest. The rule is:

   - Predict `True` for Cardiac Arrest for any value of 'Bp_change'.

4. (a) **Instance 1** (Age = 79, ECG = Hypertrophy):
      - f_1: True (Age > 56)
      - f_2: True (Hypertrophy)
      - f_3: True
      - **Aggregated**: True (Majority is True)

   (b) **Instance 2** (Age = 51, ECG = Normal):
      - f_1: False (Age ≤ 56)
      - f_2: False (Normal)
      - f_3: True
      - **Aggregated**: False (Majority is False)

   (c) **Instance 3** (Age = 48, ECG = Abnormal):
      - f_1: False (Age ≤ 56)
      - f_2: True (Abnormal)
      - f_3: True
      - **Aggregated**: True (Majority is True)

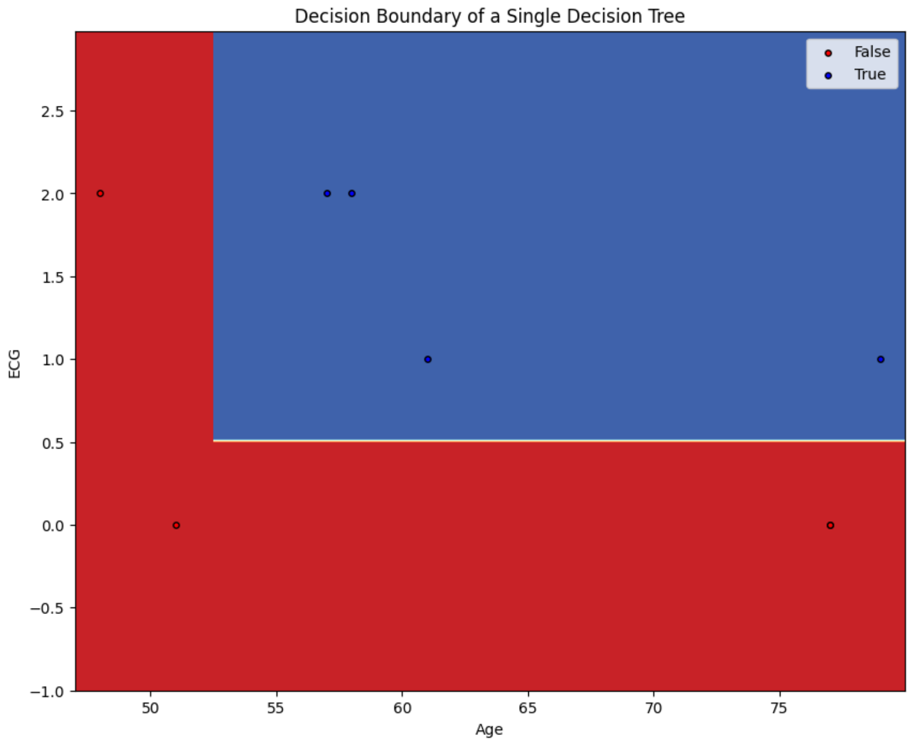
   (d) **Instance 4** (Repeated Instance 1):
      - **Aggregated**: True

(e) **Instance 5** (Age = 79, ECG = Hypertrophy):

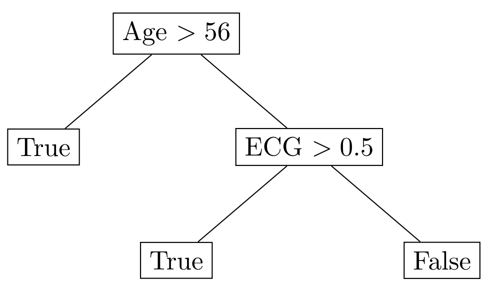
- **Aggregated**: True

(f) **Instance 6** (Age = 77, ECG = Normal):

- f_1: True (Age ¿ 56)
- f_2: False (Normal)
- f_3: True
- **Aggregated**: True (Majority is True)


Decision Boundary of a Single Decision Tree
False
True
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
ECG
50
55
60
65
70
75
Age

(b)


Age > 56
True
ECG > 0.5
True
False

Problem 7

(a) The current algorithm does not satisfy demographic parity, equality of opportunity, or individual fairness with the threshold $t = 0.5$. The positive prediction rate for Group A is 83.33% and for Group B is 50%, which are not equal, thus failing to meet demographic parity. The true positive rate for Group A is 100% and for Group B is 50%, not satisfying equality of opportunity. And, the prediction rate for outcomes 0 and 1 are 60% and 80%, respectively, indicating a potential issue with individual fairness.

(b) Finding a threshold that satisfies demographic parity or equality of opportunity without trivial decisions requires a careful balancing act, which might not be possible with a single threshold applied uniformly to all groups.

(c) Potential biases in the training process could include imbalanced training data, features correlating with group membership, or past biases reflected in the outcome variable.