CSC 311 Fall 2022 Midterm A

Thursday, October 27, 2022

Q1 [7 pts] True/False and Short-Answer Questions

Q1a (1 pt)

(True or False) The greedy algorithm for generating a decision tree is guaranteed to produce the optimal decision tree. The optimal decision tree is the smallest/most compact tree for the data set.

Solution: False

Q1b (6 pts)

Consider linear regression and logistic regression. Circle the correct answer for each statement below. If a statement is **false**, explain why in one sentence.

1. (1 pt) (True or False) They both use linear functions.

   Solution: True

2. (1 pt) (True or False) They both can be used to solve regression problems.

   Solution: False

   Logistic regression is used to solve a classification problem.

3. (1 pt) (True or False) They both use the logistic activation function.

   Solution: False

   Linear regression does not have an activation function.

Q1 [7 pts] continued

Q1c (3 pts)

Categorize each algorithm as parametric or non-parametric.
If an algorithm is parametric, describe its parameters in one sentence.

- (1 pt) K-nearest-neighbours is PARAMETRIC / NON-PARAMETRIC.

  Solution: NON-PARAMETRIC.

- (1 pt) Linear regression is PARAMETRIC / NON-PARAMETRIC.

  Solution: PARAMETRIC.

  Parameters are the weights and the bias.

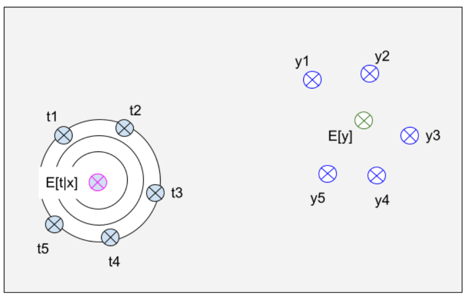
- (1 pt) The feed-forward neural network is PARAMETRIC / NON-PARAMETRIC.

  Solution: PARAMETRIC.

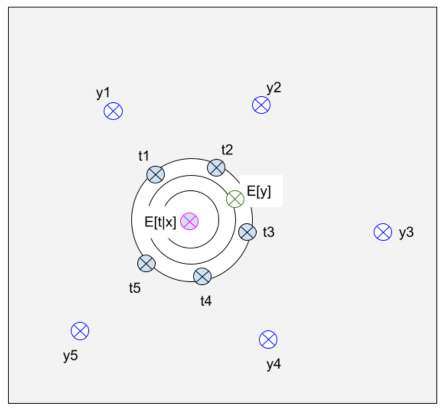
  Parameters are the weights and the bias.

Q2 [4 pts] Bias-Variance Decomposition

Figures 1a and 1b illustrate the behaviour of two machine learning algorithms. The data set contains five examples with targets $t1, \ldots, t5$. There is a fixed query point x. $E[t|x]$ is the expected target given the query point based on the true underlying distribution. Each algorithm makes five predictions: $y1, \ldots, y5$. $E[y]$ is the expected value of these predictions.



(a) Model 1



(b) Model 2

Circle the correct answer in each statement below.
Justify each answer in one sentence.

1. (2 pts) Model 1 has HIGHER / LOWER bias than Model 2.

   Justification:

   Solution:

   Model 1 has **HIGHER** bias than Model 2.

   Justification: The distance between $E[y]$ and $E[t|x]$ for model 1 is much smaller than that of model 2.

2. (2 pts) Model 1 has HIGHER / LOWER variance than Model 2.

   Justification:

   Solution:

   Model 1 has **LOWER** variance than Model 2.

Justification: The predictions for model 1 are much closer together than those for model 2.

Q3 [15 pts] Decision Trees

Consider the data set in Table 1. There are 11 examples. There are 2 binary discrete features: colour and length. "Colour" has two values: dark and light. "Length" has two values: long and short. Each example has a binary label: True or False.

Example	Colour	Length	Label
1	Dark	Long	True
2	Dark	Long	False
3	Dark	Long	False
4	Dark	Short	False
5	Dark	Short	False
6	Dark	Short	False
7	Light	Short	True
8	Light	Short	True
9	Light	Short	True
10	Light	Short	True
11	Light	Short	False

Table 1: Data Set for Decision Tree

Q3a [9 pts]

Complete the following decision tree for the data set. Note that we will split on “Colour” at the root of the tree.

- For each node, write down the number of true and false examples.
- Then, for each leaf (rectangle-shaped) node, write down the label/decision.

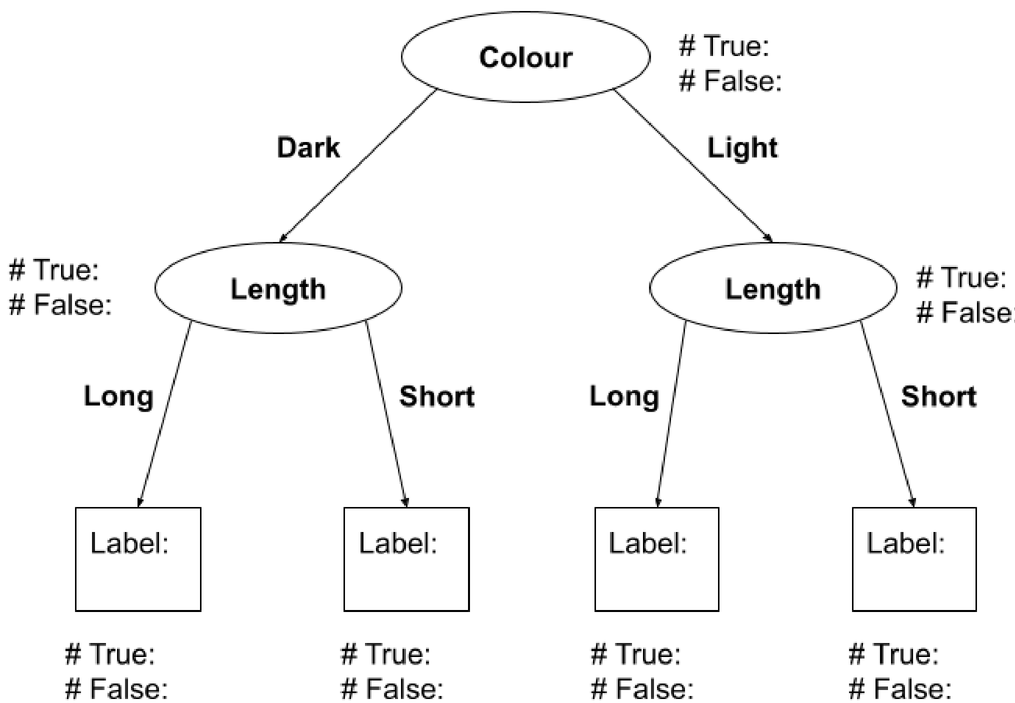


Figure 2: Complete this decision tree

Solution: The solution is in Figure 3.

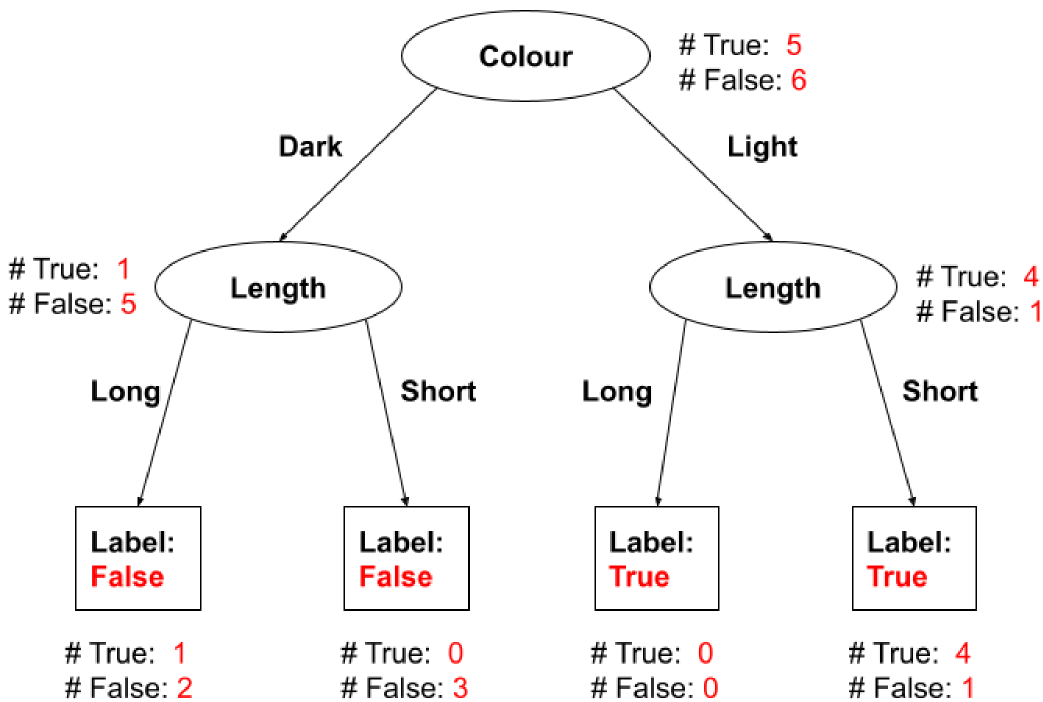


Figure 3: Decision Tree Solution A

Q3b [6 pts]

Write down the formulas for calculating the expected information gain of testing Length at the root. H denotes entropy. You do not need to calculate the result and can leave numbers as fractions. i.e. for each $\frac{\boxed{}}{\boxed{}}$ you need to enter one number in the numerator and another in the denominator.

(2 pts) The entropy before testing Length at the root

$$= H\left(\frac{\boxed{}}{\boxed{}}, \frac{\boxed{}}{\boxed{}}\right)$$

$$= -\frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}} - \frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}}$$

Q3b [6 pts] continued

(4 pts) The expected conditional entropy after testing Length at the root

$$= \frac{\boxed{}}{\boxed{}} H(\frac{\boxed{}}{\boxed{}}, \frac{\boxed{}}{\boxed{}})$$

$$+ \frac{\boxed{}}{\boxed{}} H(\frac{\boxed{}}{\boxed{}}, \frac{\boxed{}}{\boxed{}})$$

$$= \frac{\boxed{}}{\boxed{}}(-\frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}} - \frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}})$$

$$+ \frac{\boxed{}}{\boxed{}}(-\frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}} - \frac{\boxed{}}{\boxed{}} \log_2 \frac{\boxed{}}{\boxed{}})$$

Solution:

The entropy before testing Length

$$= H\left(\frac{5}{11}, \frac{6}{11}\right)$$
$$= -\frac{5}{11}\log_2 \frac{5}{11} - \frac{6}{11}\log_2 \frac{6}{11}$$

The expected conditional entropy after testing Length at the root

$$
\begin{aligned}
&= \frac{3}{11} H\left(\frac{1}{3}, \frac{2}{3}\right) + \frac{8}{11} H\left(\frac{4}{8}, \frac{4}{8}\right) \\
&= \frac{3}{11}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{8}{11}\left(-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}\right)
\end{aligned}
$$

Q4 [4 pts] Binary Linear Classification

Consider the data set in Table 2 and Figure 4. The blue circles denote positive examples and the red squares denote negative examples.

x_1	x_2	t
1	1	1
2	1	1
3	2	1
2	0	0

Table 2: Data Set for Classification

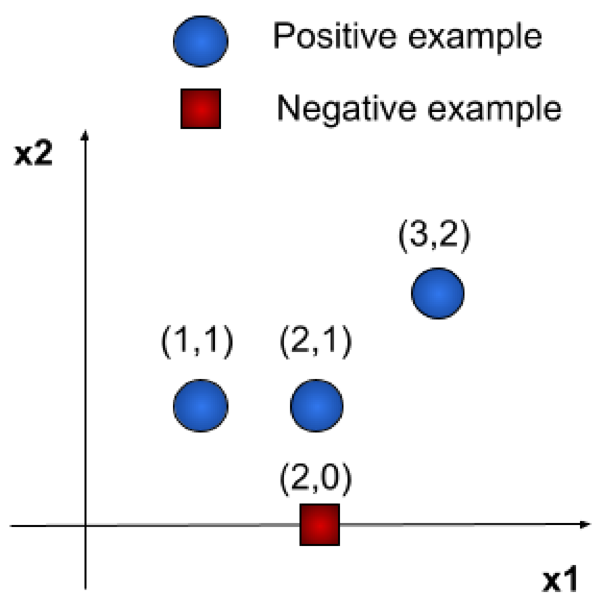


Figure 4: Data Set for Classification

Q4 [4 pts] continued

Recall the binary linear classification model with a decision rule:

$$z = w^T x$$
$$y = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0 \end{cases}$$

Assume that the bias term is zero ($b = w_0 = 0$).

Q4a (3 pts)

Write out the inequalities that represent the constraints that the weight space needs to satisfy.

Solution:

$$w_1 + w_2 \geq 0$$
$$2w_1 + w_2 \geq 0$$
$$3w_1 + 2w_2 \geq 0$$
$$2w_1 < 0$$

Q4b (1 pt)

If the problem is feasible, provide at least one assignment of the weights. Show all your work.

Solution: There are many valid solutions.

Setting $w_2 = K$ means, we need the following to hold for w_1:

$$w_{\geq} K \quad w_1 \geq \frac{K}{2} \quad w_1 \geq \frac{2K}{3}$$

which will be satisfied for all $w_1 \geq K$.

Q5 [10pts] Gradient Descent

Consider a data set with three examples. For example i, $x^{(i)}$ is the value of the input feature and $t^{(i)}$ is the value of the target.

Example	$x^{(i)}$	$t^{(i)}$
1	1	1
2	2	4
3	3	0

Table 3: Data Set for Linear Regression

We will fit a ridge regression model to the data above resulting in on weight w and one bias term b. Recall that in vector notation (with the bias term incorporated into the design matrix via concatenation into a vector of all 1s), this corresponds to training with:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}||X\mathbf{w} - \mathbf{t}|||^2 + \lambda\frac{1}{2}||\mathbf{w}||^2$$

Derive the gradient descent update rule using the parameter values below.

$$w = -1, b = 2, \alpha = 0.1, \lambda = 0.6$$

Assume that the cost function is the total loss over all the training examples. **We recommend breaking up the calculations into easy to understand steps rather than only writing down the final answer. We will award partial marks for correct intermediate steps.** You should leave your answer in the form of two equations:

$$b \leftarrow Ab + B \tag{1}$$

$$w \leftarrow Cw + D \tag{2}$$

where A, B, C, D are real valued numbers.

Solution:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, w = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, t = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$$

$$Xw - t = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -4 \\ -1 \end{bmatrix}$$

$$\frac{\partial J}{\partial w} = X^T(Xw - t) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ -4 \\ -1 \end{bmatrix} = \begin{bmatrix} -5 \\ -11 \end{bmatrix}$$

$$(1 - \alpha\lambda) = 1 - 0.1 * 0.6 = 0.94$$

$$w \leftarrow (1 - \alpha\lambda)w - \alpha\frac{\partial J}{\partial w} = 0.94w - 0.1 \begin{bmatrix} -5 \\ -11 \end{bmatrix} = 0.94w - \begin{bmatrix} -0.5 \\ -1.1 \end{bmatrix}$$

$b \leftarrow 0.94b + 0.1$

$w \leftarrow 0.94w + 1.1$