CSC 311 Fall 2022 Midterm B

Friday, October 28, 2022

Q1 Short answers [10 pts]

Q1a. (True/False) Decision Trees are a non-parametric method. [1pt]
Solution: False

Q1b. (True/False) When training a regression model, we should continue to optimize our model until the averaged training loss is exactly zero. [1pt]
Solution: False

Q1c. (True/False) Early stopping is used to prevent underfitting. [1pt]
Solution: False

Q1d. (True/False) A two-layer neural network with 2^D hidden units and a hard threshold activation function can approximate any boolean function over D input variables. [1pt]
Solution: True

Q1e. You have trained a logistic regression model on a very large dataset up to an accuracy of 77%.(a) Your colleague claims that by using a neural network you can outperform the Bayes Optimal classifier – are they correct? (Yes/No with short justification) (b) You now train a neural network to do binary classification via the backpropagation algorithm on the dataset - do you expect your accuracy to increase/stay the same or decrease? Provide a short justification. [2pt]

Solution: (a) No; Bayes error is irreducible. (b) Increase - neural network represents a larger hypothesis class.

Q1f. Consider the following dataset representing the XNOR boolean function.

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	1

Is the data linearly separable? Briefly justify your choice on whether the data is (or is not) linearly separable. [2pts]

Not linearly separable. Justification – can prove via contradiction using argument in class. Alternatively, notice that this is simply XOR with the labels flipped, so if XOR is not linearly separable, then neither is XNOR.

Q1g. Why can a neural network with the hard-threshold activation function correctly classify the XOR function as we saw in lecture? Your answer should refer to the hypothesis space of functions represented by each model. Highlight one technical challenge that can arise when using backpropagation to learn neural networks with a hard threshold activation function. [2pts]

Solution: (a) Larger hypothesis space in neural network; learns a hidden representation wherein XOR is linear in the hidden representation. (b) The discontinuity at 0 means gradients wrt inputs do not exist.

Q2 [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	0
2	2	4
3	3	-1

Table 1: 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:

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

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

$$w = 2, b = -1, \alpha = 0.2, \lambda = 0.5$$

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:

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

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

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

Solution:

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

$$Xw - t = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0 \\ 4 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 6 \end{bmatrix}$$

$$\frac{\partial J}{\partial w} = X^T(Xw - t) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 6 \end{bmatrix} = \begin{bmatrix} 6 \\ 19 \end{bmatrix}$$

$$(1 - \alpha\lambda) = 1 - 0.2 * 0.5 = 0.9$$

$$w \leftarrow (1 - \alpha\lambda)w - \alpha\frac{\partial J}{\partial w} = 0.9w - 0.1 \begin{bmatrix} 6 \\ 19 \end{bmatrix} = 0.9w - \begin{bmatrix} 0.6 \\ 1.9 \end{bmatrix}$$

$$b \leftarrow 0.9b - 0.6$$

$$w \leftarrow 0.9w - 1.9$$

Q3 Feature maps and separability [5pts]

Q3a. [1pts] In lecture, we discussed polynomial regression in one variable. It is also possible to fit polynomials of two variables. A degree-2 polynomial in two features x_1 and x_2 is a sum of monomials which are at most quadratic in x_1 and x_2. One such example is: $x_1^2 - 3x_1, x_2 + 2x_2 + 12$. Design a feature map $\phi(u, v)$ that allows you to use a linear regression solver to fit *any* degree-2 polynomials. No justification is required.

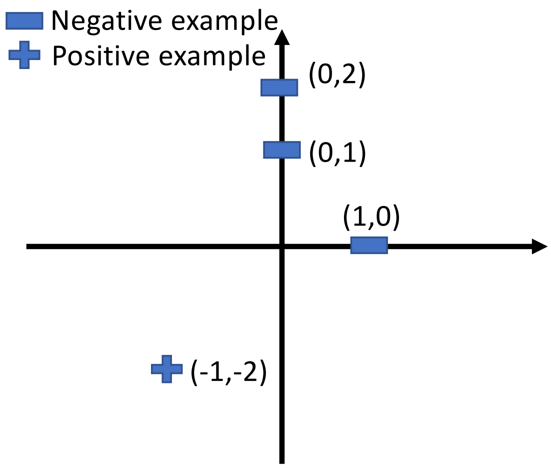

Figure 1: Binary Dataset

Solution: $\phi(u, v) = [1, x_1, x_2, x_1x_2, x_1^2, x_2^2]$

Q3b. [4pts] Figure 1 presents a visual depiction of a binary dataset. 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 ($b = w_0 = 0$).

Q3b Answer

1. Write out the inequalities that represent the constraints that the weight space needs to satisfy. [3pts]
2. If the problem is feasible, provide one assignment of the weights. [1pts]

Solution: (1) $w_1 < 1, 2w_2 < 0, w_2 < 0, -w_1 - 2w_2 \geq 0$ (2) Yes (3) $w_1 = 0.5, w_2 = -1$

Q4 [15pts] Decision Trees

Consider the data set in Table 2. 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	True
3	Dark	Long	False
4	Dark	Short	True
5	Dark	Short	True
6	Dark	Short	True
7	Light	Short	False
8	Light	Short	False
9	Light	Short	False
10	Light	Short	False
11	Light	Short	True

Table 2: Data Set for Decision Tree

Q4a. [9pts] Complete the following decision tree for the data set.

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

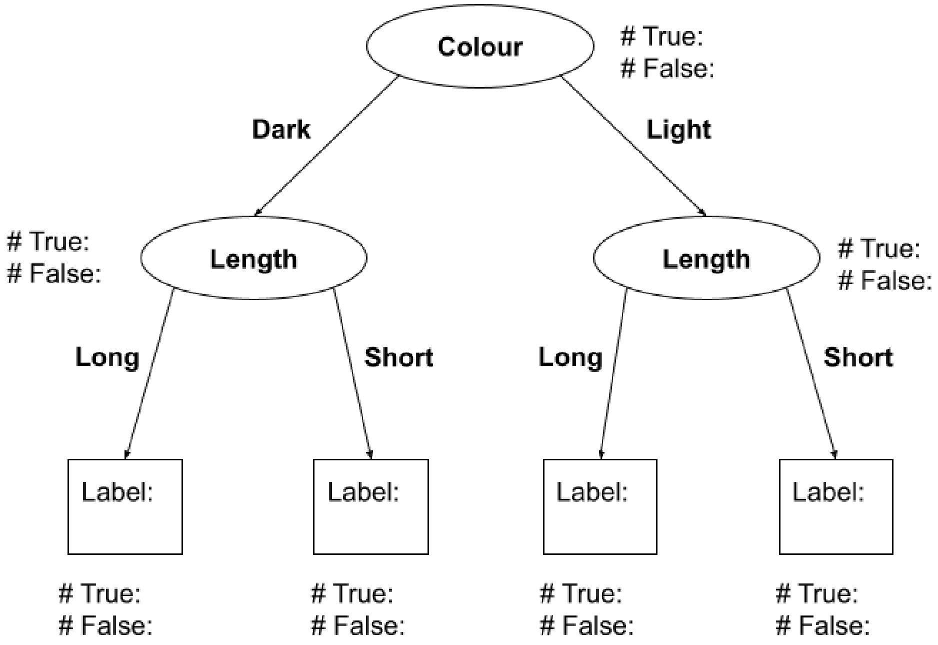

Figure 2: Complete this decision tree

Solution: The solution is in Figure 3.

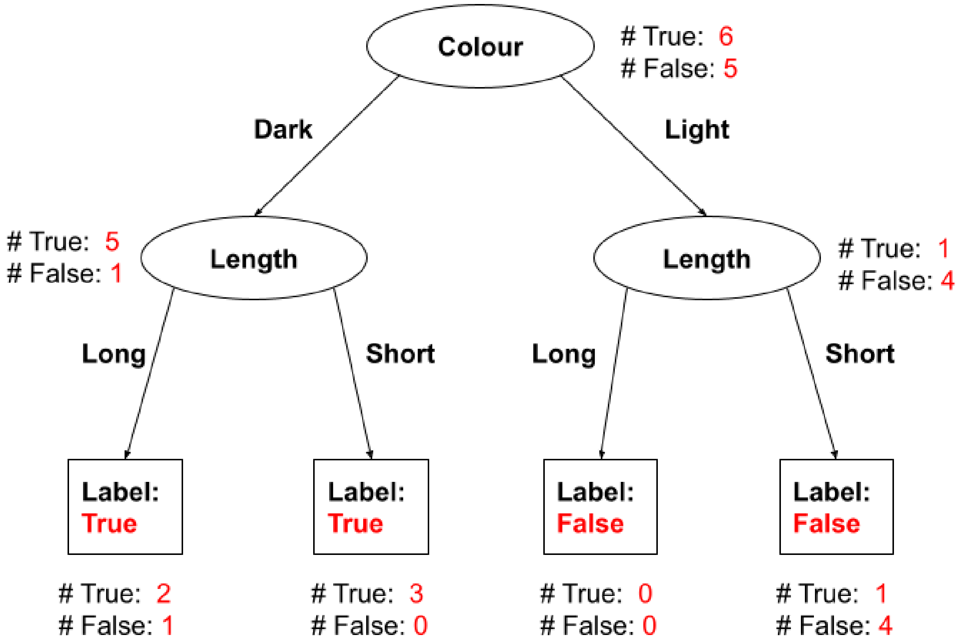

Figure 3: Decision Tree Solution B

Q4b. [6pts] 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 $\frac{\Box}{\Box}$ you need to enter one number in the numerator and another in the denominator.

(2 pts) The entropy before testing Length

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

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

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

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

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

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

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

Solution: The entropy before testing Length

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

The expected conditional entropy after testing Length at the root

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