

## CIS 419/519: Quiz 2

September 20, 2019

1. We have an attribute set made of two boolean features  $(A, B)$  where  $A, B \in \{0, 1\}$  with boolean labels  $y \in \{0, 1\}$ . In our dataset we have 4 categories of data points of  $\langle x, y \rangle$  with counts:

- (i)  $\langle (A = 0, B = 1), 1 \rangle$ : 25 examples
- (ii)  $\langle (A = 0, B = 0), 0 \rangle$ : 0 examples
- (iii)  $\langle (A = 1, B = 0), 1 \rangle$ : 10 examples
- (iv)  $\langle (A = 1, B = 1), 0 \rangle$ : 50 examples

Say we want to train a decision tree with this data. Which feature should we split on first, and what is the information gain?

- (a) A : IG=0.519
- (b) B : IG=0.167
- (c) A : IG=0.650
- (d) B : IG=0.059

2. The Boolean function,  $x_1 \wedge x_3 \wedge x_4$ , is a linear function over the boolean variables,  $x_1, x_2, x_3, x_4$ . Which of the following is a correct “linear” representation for it?

- (a)  $x_1 + x_3 + x_4 \geq 0$
- (b)  $x_1 + x_3 + x_4 \geq 3$
- (c)  $x_1 + x_2 + x_3 + x_4 \geq 0$
- (d)  $x_1 + x_3 + x_4 \leq 3$

3. We want to show that the Boolean function ‘y = 1 if and only if at least 6 out of 10 variables are 1’ can be written as a linear threshold function  $w^T \cdot x \geq \theta$ . Here  $x \in \{0, 1\}^{10}$ , and  $y \in \{0, 1\}$ . What  $w^T, \theta$  will show this?

- (a)  $w^T = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]$ ,  $\theta = 1$
- (b)  $w^T = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$ ,  $\theta = 6$
- (c)  $w^T = [0, 0, 0, 0, 1, 1, 1, 1, 1, 1]$ ,  $\theta = 6$

(d)  $w^T = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1], \theta = 1$

4. In a machine learning task, we want to define features for a classifier where the input consists of a pair of character strings composed of lower case letters  $\{a, b, c, \dots, z\}$ . For example 'four seasons' or 'bill clinton' are possible input strings. We define two feature types:
- (i) "The last character in the first string is \_" (Note that this will result in a boolean feature)
  - (ii) "The first character in the second string is \_" (Note that this will result in a boolean feature)

If we generate a feature space using only these two types, the dimensionality of the feature space will then be:

- (a) 26
  - (b) 2
  - (c) 52
  - (d) 54
5. In a machine learning task, we want to define features for a classifier where the input consists of a pair of character strings composed of lower case letters  $\{a, b, c, \dots, z\}$ . For example 'four seasons' or 'bill clinton' are possible input strings. We define two feature types:
- (i) "Whether or not the first character in the first string is a vowel" (Note that this will result in a boolean feature)
  - (ii) "The first character in the second string is \_" (Note that this will result in a boolean feature)

If features in each example are sorted by alphabetic order (ie. features for 'a' come before those for 'b'), and the features corresponding to type (i) appear before those that correspond to type (ii), which of the following examples is the feature based representation of 'vijay kumar'?

- (a) 1 ; 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0
- (b) 0 ; 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1
- (c) 0 ; 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
- (d) 1, 1