CSCI567 Theory Assignment 1

Yifan Wang

3038184983

wang608@usc.edu

Problem 1 Decision Tree (8 points)

In this problem, you are given four 2-dimensional data points as shown in Table 1:

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

Table 1: Four 2-dimensional data points and their labels.

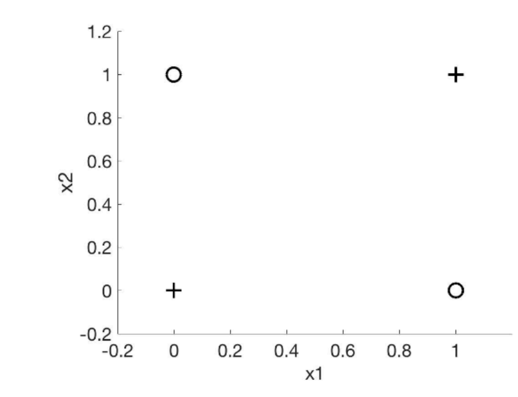

Figure 1: The plus sign means label $y = 0$ and the circle means have $y = 1$.

1.1 Fig. 2 is a decision tree of the given data with zero training error.

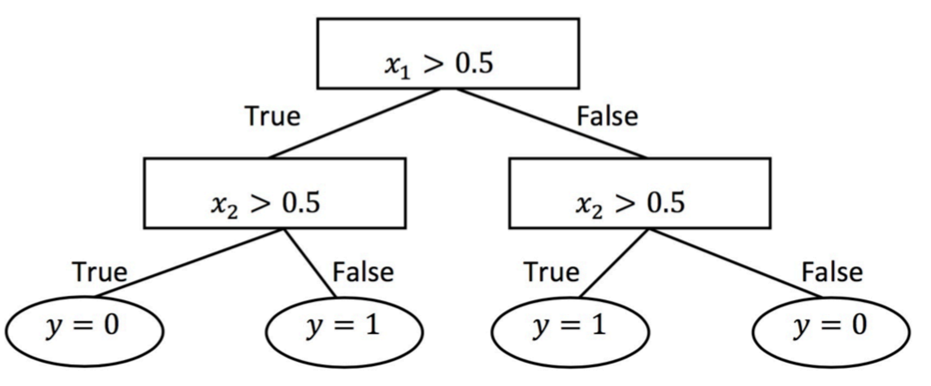

Figure 2: The decision tree with zero training error.

Suppose now you have two test data points:

x_1	x_2	label y
0.8	0.8	0
0.6	0.4	1

What would be your test error based on decision tree in Fig. 2? (Define the test error as the fraction of mis-classifications made on the testing set.) **(2 points)**

$(0.8, 0.8)$ the tree gives prediction 0

$(0.6, 0.4)$ the tree gives prediction 1

The test error would be 0

1.2 Now consider a new decision tree in Fig. 3. Note that the depth of the new decision tree is 1, and it does not have zero training error for the given data anymore.

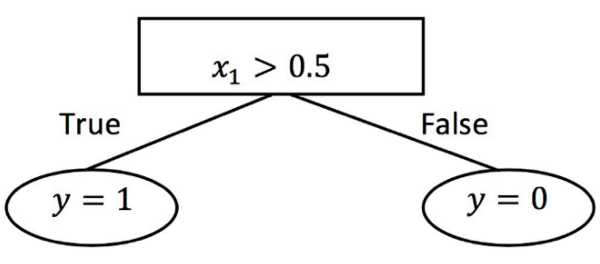

Figure 3: The decision tree with depth = 1.

Given the two test data points from Question 1.1, what would be the test error using the new decision tree? **(2 points)**

$(0.8, 0.8)$ the tree gives prediction 1

$(0.6, 0.4)$ the tree gives prediction 1

The test error would be 0.5

1.3 Is the decision tree in Fig. 3 a linear or non-linear classifier in terms of (x_1, x_2) (Yes/No)? Can you classify the given data in Table 1 and get zero classification error by drawing a depth-1 decision tree similar to Fig. 3 (Yes/No)? Note that the decision rule should be based on a single variable (x_1 or x_2) in the rectangle (e.g., $x_1 > 1$ or $x_2 < 2$). **(2 points)**

It is a non-linear classifier

It is impossible to get a single variable based decision tree to get zero classification error

1.4 If you can put any expression of variables in the rectangle (e.g.: $f(x_1, x_2) \geq c$ or $f(x_1, x_2) < c$, for any function f and real number c), can you classify the given data in Table 1 and get zero classification error by drawing a depth-1 decision tree similar to Fig. 3 (Yes/No)? Please also **briefly** justify your answer. **(2 points)**

Yes

for example

$$y = 0, \ if \ |x_1 - x_2| \leq 0.5$$
$$y = 1, \ else,$$

Problem 2 kNN classification (9 points)

Let n be the number of training points, and each point $\mathbf{x} \in \mathbb{R}^d$ haslabel $y \in \{0, 1\}$ drawn from $P(y|\mathbf{x})$. Assume that any label y depends only on a correspondent training point $\mathbf{x}$ and does not depent on anything else (i.e. $P(y|\mathbf{x}, \mathbf{x}') = P(y|\mathbf{x})$ for any $\mathbf{x}'$).

2.1 Show that for any $\mathbf{x}$ and $\mathbf{x}'$ with labels y and y' respectively

$$P(y, y'|\mathbf{x}, \mathbf{x}') = P(y|\mathbf{x})P(y'|\mathbf{x}')$$

using the label independence property above. **(2 points)**

$$P(y, y'|x, x') = P(y|x, x')P(y'|x, x') = P(y|x)P(y'|x')$$

2.2 For a test point $\mathbf{x}_t$ with label y^*, its nearest neighbor $\mathbf{x}_{\text{NN}}$ has label y. Assume that, as $n \to \infty$, $\mathbf{x}_{\text{NN}}$ satisfies $||\mathbf{x}_{\text{NN}} - \mathbf{x}_t|| \to \mathbf{0}$ with probability 1. Show that, when $n \to \infty$

$$P(y^* \neq y|\mathbf{x}_{\text{NN}}, \mathbf{x}_t) \to 2P(y^* = 0|\mathbf{x}_t)P(y^* = 1|\mathbf{x}_t)$$

with probability 1. **(3 points)**

$$P(y^* \neq y|x_{NN}, x_t) = P(y^* = 0, y = 1|x_{NN}, x_t) + P(y^* = 1, y = 0|x_{NN}, x_t)$$
$$= P(y^* = 0|x_t)P(y = 1|x_{NN}) + P(y^* = 1|x_t)P(y = 0|x_{NN}) \quad (1)$$

when $n \to \infty$

$$||x_{NN} - x_t|| \to 0$$

which implies

$$P(y = y^*| \, ||x_{NN} - x_t|| \to 0) = 1$$

then Eq.(1) becomes,

$$P(y = 0|x_{NN})P(y^* = 1|x_t) + P(y = 1|x_{NN})P(y^* = 0|x_t)$$
$$= 2P(y^* = 0|x_t)P(y^* = 1|x_t)$$

2.3 Prove the inequality

$$P(y \neq y^*|\mathbf{x}_{\text{NN}}, \mathbf{x}_t) \leq \min_y 2P(y|\mathbf{x}_t)$$

based on the result above. **(2 points)**

$$P(y^* = 0|x_t)P(y = 1|x_{NN}) + P(y^* = 1|x_t)P(y = 0|x_{NN}) \leq 2P(y^* = 0|x_t)P(y = 1|x_{NN}) \leq min_y 2P(y|x_t)$$

2.4 Recall that the Bayes optimal classifier predicts 1 if $P(y = 1|\mathbf{x}) > 0.5$ and 0 otherwise. What does $\min_y P(y|\mathbf{x})$ mean for the optimal classifier? What does this result tell us about NN classification, in terms of the optimal classifier? **(2 points)**

$min_y P(y|x)$ means the error rate/risk given an input x

NN classification is a Bayes optimal classifier

Problem 3 Boosting (AdaBoost) (15 points)

Two rounds of boosting

3.1 Two rounds of boosting You have six training points (A, B, C, D, E, F) and five classifiers (h1, h2, h3, h4, h5) which make the following misclassifications in Table 3.1. **(10 points)**

Classifier	Misclassified training points (A, B, C, D, E, F)					
h1	A			D		F
h2				D		
h3		B	C			
h4	A	B				F
h5		B	C	D		

Perform two rounds of boosting with these classifiers and training data. In each round, pick the classifier with the **lowest error rate**. Break ties by picking the classifier that comes first in this list: h1, h2, h3, h4, h5.

```
----------------------------------------
                  | Round 1 | Round 2
----------------------------------------
weight-A          | 1/6     | 1/10
weight-B          | 1/6     | 1/10
weight-C          | 1/6     | 1/10
weight-D          | 1/6     | 5/10
weight-E          | 1/6     | 1/10
weight-F          | 1/6     | 1/10
----------------------------------------
Error Rate h1     | 3/6     | 7/10
Error Rate h2     | 1/6     | 5/10
Error Rate h3     | 2/6     | 2/10
Error Rate h4     | 3/6     | 3/10
Error Rate h5     | 3/6     | 7/10
----------------------------------------
weak classifier   | h2      | h3
classifier error  | 1/6     | 2/10
voting power      | 0.8     | 0.6931
----------------------------------------
```

3.2 Three of the training points (B, D, F) have been selected below. For each one, decide whether the ensemble classifier $H(x)$ produced after two rounds of boosting misclassifies or correctly classifies that point. Circle the best answer in each case. If the answer can't be determined from the available information, circle ?Can't tell? instead. **(3 points)**

Training point	Classification by ensemble classifier		
B	Correctly classified	Misclassified	Can't tell
D	Correctly classified	Misclassified	Can't tell
F	Correctly classified	Misclassified	Can't tell

point B: sign(0.81+0.69-1) = correctly classified

point D: sign(0.8-1+0.691) = misclassified

point F: sign(0.81+0.691) = correctly classified

3.3 Suppose you continue the AdaBoost procedure from Part A for a total of 2021 rounds. (You may assume it doesn't terminate before then.) If you always pick the classifier with the lowest error rate, which training data point will have the smallest weight at the end of the 2021st round? Choose one of A,B,C,D,E,F. Please give a brief reason for your choice. **(2 points)**

E will have the smallest weight, since all the weak classifier predicts correctly on it.