ZIYAO YANG 9490544195

Theory Assignment 1

1. if $C(\vec{x_i}, \vec{x_j}) \leq C(\vec{x_i}, \vec{x_0})$

$$\Rightarrow 1 - \frac{\sum_{d=1}^{D}(x_{id} \cdot x_{jd})}{\|\vec{x_i}\|_2 \cdot \|\vec{x_j}\|_2} \leq 1 - \frac{\sum_{d=1}^{D} C(x_{id}, x_{od})}{\|\vec{x_i}\|_2 \cdot \|\vec{x_0}\|_2}$$

since $\|\vec{x_i}\|_2 = \|\vec{x_j}\|_2 = \|\vec{x_0}\|_2 = 1$

then $\sum_{d=1}^{D}(-x_{id} \cdot x_{jd}) \leq \sum_{d=1}^{D}(-x_{id}, x_{od})$

$$\Rightarrow \sum_{d=1}^{D} -2x_{id} \cdot x_{jd} \leq \sum_{d=1}^{D}(-2x_{id} \cdot x_{od})$$

$$\Rightarrow 2 - \sum_{d=1}^{D} x_{id} \cdot x_{jd} \leq 2 - \sum_{d=1}^{D}(2x_{id} x_{od})$$

$$\Rightarrow \sum_{d=1}^{D}(x_{id})^2 + \sum_{d=1}^{D}(x_{jd})^2 - \sum_{d=1}^{D} 2x_{id} \cdot x_{jd} \leq$$

$$\sum_{d=1}^{D}(x_{id})^2 + \sum_{d=1}^{D}(x_{od})^2 - \sum_{d=1}^{D} 2x_{id} \cdot x_{od}$$

$$\Rightarrow \sum_{d=1}^{D}(x_{id} - x_{jd})^2 \leq \sum_{d=1}^{D}(x_{id} - x_{od})^2$$

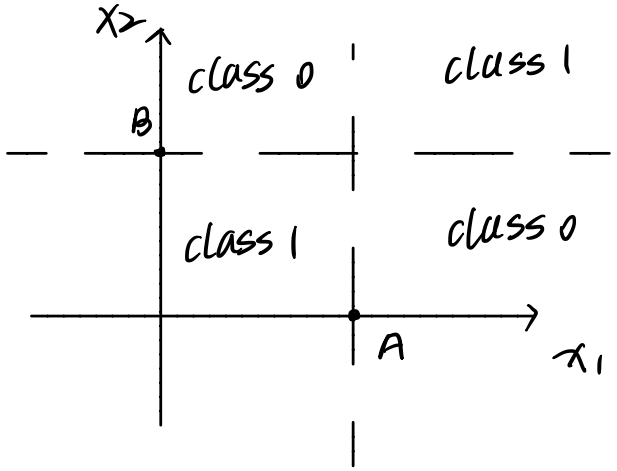
$$\Rightarrow E(\vec{x_i}, \vec{x_j}) \leq E(\vec{x_i}, \vec{x_0})$$

2. Since 100 dimensional vector are distinct and don't repeat themselves in the dataset, each $\vec{x}$ only has one label. And each split only could generate 2 branches, so that the decision Tree is binary Tree, which could include all the paths. when it is completely binary Tree.

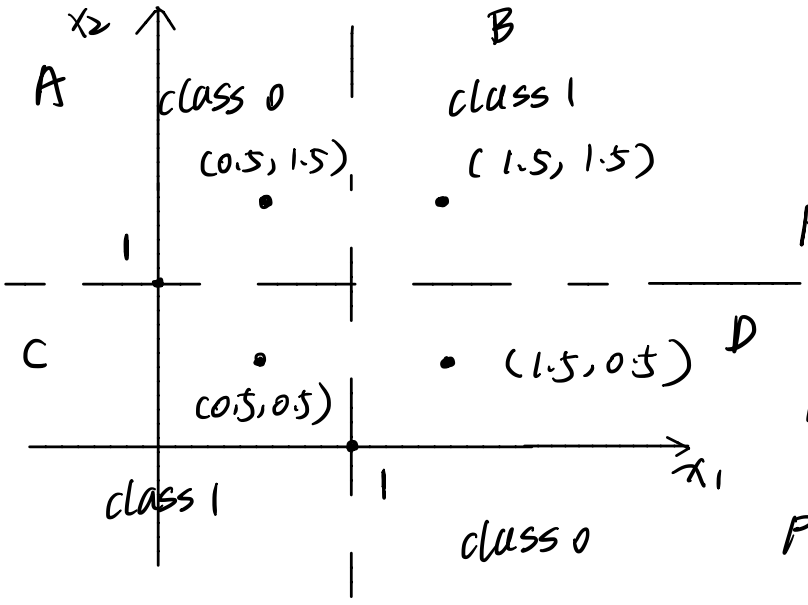
Therefore, no mater where $\vec{x}$ is, it always could find its unique table, which means the Decision Tree could classify the dataset with No error.

Yes, we could specify 1-NN to result in exactly the same classification as our decision Tree. Because $\vec{x}$ are distinct, which means all the points are distinct and 1-NN could find the nearest one to specify its label.

3. From the Decision Tree, we could get:



suppose A=B=1, so we could use 4 points to train KNN.



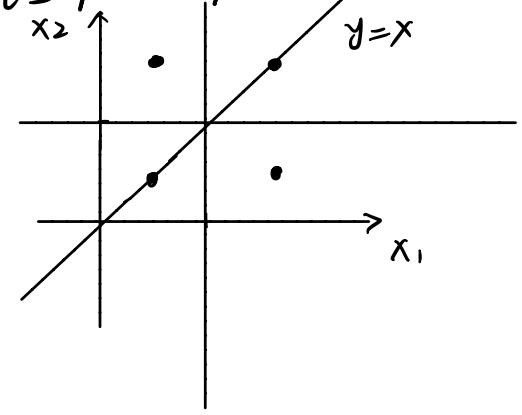
points in field A are nearest to (0.5, 1.5)

points in B are nearest to (1.5, 1.5)

points in C are nearest to (0.5, 0.5)

points in D are nearest to (1.5, 0.5)

Let's prove points in A is nearest to (0.5, 1.5)



since $x_1=1$ is equidistant line from (0.5, 1.5) and (1.5, 1.5)

$x_2=1$ is equidistant line from (0.5, 1.5) and (0.5, 0.5)

$y=x$ is equidistant line from (0.5, 1.5) and (1.5, 0.5)

So points in field A: $\{ x_1<1,\ x_2>1,\ y<x \}$ is nearest to (0.5, 1.5)

4.1

x_1	x_2	label y	predicted label
0.2	0.2	0	0
0.2	0.8	1	1

test error $= \frac{0}{2} = 0$

4.2

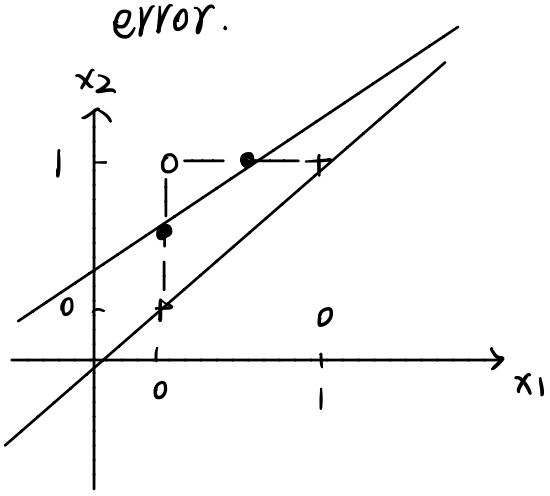
x_1	x_2	label y	predicted label
0.2	0.2	0	0
0.2	0.8	1	0

test error $= \frac{1}{2} = 0.5$

4.3 Yes, the decision Tree in figure 4 is linear classification based on $X_1 = 0.5$.

No, depth-1 decision Tree cannot classify data without error.

4.4 No. I cannot classify the data with combination of x_1 and x_2. since there doesn't exist a descision boundary, depth-1 decision tree cannot get zero classification error.



If you wanna to split (0,1) from (0,0) and (1,1), you must choose 2 points from 2 imaginary line respectively. However, (1,0) is always opposite to (0,1).
Therefore you cannot find a line to split completely

5.1

	T_1		T_2	
	left child	right child	left child	right child
examples	150 A 50 B	50 A 150 B	0 A 100 B	200 A 100 B
classify error	$\frac{50}{200} = 0.25$	$\frac{50}{200} = 0.25$	0	$\frac{100}{300} = 0.33$
total error	$\frac{100}{400} = 0.25$		$\frac{100}{400} = 0.25$	
child entropy	$-\frac{1}{4}\ln\frac{1}{4} - \frac{3}{4}\ln\frac{3}{4}$ ≈ 0.56	$-\frac{1}{4}\ln\frac{1}{4} - \frac{3}{4}\ln\frac{3}{4}$ ≈ 0.56	0	$-\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3}$ ≈ 0.64
Gini Impurity	$1-(\frac{1}{10}+\frac{9}{10})$ ≈ 0.38	$1-(\frac{1}{10}+\frac{9}{10})$ ≈ 0.38	$1-(0+1)$ $= 0$	$1-(\frac{1}{4}+\frac{4}{4})$ ≈ 0.44

5.2

	T_1	T_2
classification error	$\frac{100}{400} = 0.25$	$\frac{100}{400} = 0.25$
conditional entropy	$\frac{1}{2}\times 0.56 + \frac{1}{2}\times 0.56 = 0.56$	$\frac{1}{4}\times 0 + \frac{3}{4}\times 0.64 = 0.48$
weighted Gini impurity	$\frac{1}{2}\times 0.38 + \frac{1}{2}\times 0.38 = 0.38$	$\frac{1}{4}\times 0 + \frac{3}{4}\times 0.44 = 0.33$

Conditional entrophy(T_2) is smaller means Information Gain is bigger so T_2 is better. weighted Gini impurity (T_2) is smaller, which means it has better ordering after spliting. T_2 is better.

6.1 $P(\text{play Tennis} = \text{Yes}) = \frac{2}{3}$

$P(\text{play Tennis} = \text{No}) = \frac{1}{3}$

6.2 $P(\text{Weather} = \text{sunny} \mid \text{play Tennis} = \text{Yes})$

$$= \frac{P(\text{weather} = \text{sunny}, \text{play Tennis} = \text{Yes})}{P(\text{play Tennis} = \text{Yes})} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

$P(\text{Emotion} = \text{Normal} \mid \text{play Tennis} = \text{Yes}) = \frac{\frac{1}{6}}{\frac{2}{3}} = \frac{1}{4}$

$P(\text{Homework} = \text{much} \mid \text{play Tennis} = \text{Yes}) = \frac{\frac{1}{6}}{\frac{2}{3}} = \frac{1}{4}$

6.3 $P(\text{play Tennis} = \text{Yes} \mid \vec{x}) = \frac{P(\vec{x} \mid \text{play Tennis} = \text{Yes}) \cdot P(\text{play Tennis} = \text{Yes})}{P(\vec{x})}$

$P(\vec{x} \mid \text{play Tennis} = \text{Yes}) = \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{32}$

Similarly

$P(\vec{x} \mid \text{play Tennis} = \text{No}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

$P(\text{play Tennis} = \text{Yes}) = \frac{2}{3}$, $P(\text{play Tennis} = \text{No}) = \frac{1}{3}$

$P(\vec{x}) = P(\vec{x} \mid \text{play Tennis} = \text{Yes}) \cdot P(\text{play Tennis} = \text{Yes}) +$

$P(\vec{x} \mid \text{play Tennis} = \text{No}) \cdot P(\text{play Tennis} = \text{No})$

$= \frac{1}{32} \times \frac{2}{3} + \frac{1}{8} \times \frac{1}{3} = \frac{1}{16}$

$P(\text{play Tennis} = \text{Yes} \mid \vec{x}) = \frac{\frac{1}{32} \times \frac{2}{3}}{\frac{1}{16}} = \frac{1}{3}$

$P(\text{play Tennis} = \text{No} \mid \vec{x}) = 1 - \frac{1}{3} = \frac{2}{3}$