Exercise # of Machine Learning [IN 2064]

Name: Yiman Li
Matr-Nr: 03724352

Problem 1

The $L_1 - norm$ and $L_2 - norm$ are shown in Table 1 and Table 2. With respect to question c), comparing the two distance measures and correspongding classification result (especially see the result of point D), we can see that the classification result strongly depends on the way how the diatance between 2 points are defined.

Problem 2

a) A new point under unweighted classifier will be absolutely distribued to the dataset C since k is the number of all data points, and according to the unweighted classification tasks which is described as $\hat{y} = arg\ max\ \frac{1}{k} \sum \mathrm{II}(y_i = c)$, this point will be labeled with the majority of the whole label set, which is C.

b) For a new point under weighted (by distance) version of k-Nearest Neighbors, which is in the form that $\hat{y} = arg\ max\ \frac{1}{Z} \sum \frac{1}{d(\boldsymbol{x}, \boldsymbol{x_i})} \mathrm{II}(y_i = c)$, since we don't know the expression of the distance $d(\boldsymbol{x}, \boldsymbol{x_i})$ here, so we can ensure the result of our classification.

Problem 3

For a standard decision tree of depth 1, depth = 1, which means we can choose only one splitting variable, under which we can only draw a line that is parallel to the x_1 or x_2 axis, so we can not find such a tree to classify this dataset with 100% accuracy while the only solution is $x_1 < x_2$.

Problem 4

a) Using the definition of entropy, we can figure out that:

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0.4 \times log_2 0.4 + 0.6 \times \log_2 0.6) = 0.971$$

b) Since that

$$Gain(D, x) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{D} \mathrm{Ent}(D^v)$$

- When we choose x_1 as the splitting variable, then

$$\mathrm{Ent}(D_1^1) = -(\frac{2}{5} \cdot \log_2 \frac{2}{5} + \frac{3}{5} \cdot \log_2 \frac{3}{5}) = 0.971$$
$$\mathrm{Ent}(D_1^2) = -(\frac{2}{5} \cdot \log_2 \frac{2}{5} + \frac{3}{5} \cdot \log_2 \frac{3}{5}) = 0.971$$
$$Gain(D, x_1) = 0.97095 - \frac{5}{10} \times 0.971 - \frac{5}{10} \times 0.971 = 0$$

Table 1: L_1 norm

P_1	P_2	$\|\triangle x\|$	$\|\triangle y\|$	$norm$	1-NN
A	B	1.0	0.5	1.5	B/C
	C	0.0	1.5	1.5	
	D	2.0	2.5	4.5	
	E	4.5	2.5	7.0	
	F	4.5	1.5	6.0	
B	A	1.0	0.5	1.5	A
	C	1.0	2.0	3.0	
	D	1.0	3.0	4.0	
	E	3.5	3.0	6.5	
	F	3.5	2.0	5.5	
C	A	0.0	1.5	1.5	A
	B	1.0	2.0	3.0	
	D	2.0	1.0	3.0	
	E	4.5	1.0	5.5	
	F	4.5	0.0	4.5	
D	A	2.0	2.5	4.5	E
	B	1.0	3.0	4.0	
	C	2.0	1.0	3.0	
	E	2.5	0.0	2.5	
	F	2.5	1.0	3.5	
E	A	4.5	2.5	7.0	F
	B	3.5	3.0	6.5	
	C	4.5	1.0	5.5	
	D	2.5	0.0	2.5	
	F	0.0	1.0	1.0	
F	A	4.5	1.5	6.0	E
	B	3.5	2.0	5.5	
	C	4.5	0.0	4.5	
	D	2.5	1.0	3.5	
	E	0.0	1.0	1.0	

Table 2: L_2 norm

P_1	P_2	$\|\triangle x\|^2$	$\|\triangle y\|^2$	$norm$	1-NN
A	B	1.00	0.25	$\sqrt{1.25}$	B
	C	0.00	2.25	$\sqrt{2.25}$	
	D	4.00	6.25	$\sqrt{10.25}$	
	E	20.25	6.25	$\sqrt{26.50}$	
	F	20.25	2.25	$\sqrt{22.50}$	
B	A	1.00	0.25	$\sqrt{1.25}$	A
	C	1.00	4.00	$\sqrt{5.00}$	
	D	1.00	9.00	$\sqrt{10.00}$	
	E	12.25	9.00	$\sqrt{21.25}$	
	F	12.25	4.00	$\sqrt{16.25}$	
C	A	0.00	2.25	$\sqrt{2.25}$	A
	B	1.00	4.00	$\sqrt{5.00}$	
	D	4.00	1.00	$\sqrt{5.00}$	
	E	20.25	1.00	$\sqrt{21.25}$	
	F	20.25	0.00	$\sqrt{20.25}$	
D	A	4.00	6.25	$\sqrt{10.25}$	C
	B	1.00	9.00	$\sqrt{10.00}$	
	C	4.00	1.00	$\sqrt{5.00}$	
	E	6.25	0.00	$\sqrt{6.25}$	
	F	6.25	1.00	$\sqrt{7.25}$	
E	A	20.25	6.25	$\sqrt{26.50}$	F
	B	12.25	9.00	$\sqrt{21.25}$	
	C	20.25	1.00	$\sqrt{21.25}$	
	D	6.25	0.00	$\sqrt{6.25}$	
	F	0.00	1.00	$\sqrt{1.00}$	
F	A	20.25	2.25	$\sqrt{22.50}$	E
	B	12.25	4.00	$\sqrt{16.25}$	
	C	20.25	0.00	$\sqrt{20.25}$	
	D	6.25	1.00	$\sqrt{7.25}$	
	E	0.00	1.00	$\sqrt{1.00}$	

- When we choose x_2 as the splitting variable, then

$$\text{Ent}(D_2^1) = -(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4}) = 1$$

$$\text{Ent}(D_2^2) = -(\frac{2}{6} \cdot \log_2 \frac{2}{6} + \frac{4}{6} \cdot \log_2 \frac{4}{6}) = 0.918$$

$$Gain(D, x_2) = 0.971 - \frac{4}{10} \times 0.971 - \frac{6}{10} \times 0.918 = 0.012$$

- When we choose x_3 as the splitting variable, then

$$\text{Ent}(D_3^1) = -(\frac{3}{5} \cdot \log_2 \frac{3}{5} + \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0.971)$$

$$\text{Ent}(D_3^2) = -(\frac{1}{5} \cdot \log_2 \frac{1}{5} + \frac{4}{5} \cdot \log_2 \frac{4}{5} = 0.722)$$

$$Gain(D, x_3) = 0.97095 - \frac{5}{10} \times 0.971 - \frac{5}{10} \times 0.722 = 0.125$$

Since that $Gain(D, x_3) > Gain(D, x_2) > Gain(D, x_1)$, so we choose x_3 as the splitting variable, and the data 1,2,3,7,8 will be classified in to the catagory of "Skill", and the data 4,5,6,9,10 will be classified into the catagory of "Chance".

Probelm 5

References