

1. For the following dataset:

apple	ibm	lemon	sun	CLASS
TRAINING INSTANCES				
4	0	1	1	FRUIT
5	0	5	2	FRUIT
2	5	0	0	COMPUTER
1	2	1	7	COMPUTER
TEST INSTANCES				
2	0	3	1	?
1	2	1	0	?

(i). Using the Euclidean distance measure, classify the test instances using the 1-NN method.

Euclidean (2-norm) : vectors A & B

$$d_E(A, B) = \sqrt{\sum_k (a_k - b_k)^2}$$

1-NN for  $T_1 : \langle 2, 0, 3, 1 \rangle$

$$\begin{aligned} d_E(T_1, A) &= \sqrt{(4-2)^2 + (0-0)^2 + (1-3)^2 + (1-1)^2} \\ &= \sqrt{4+0+4+0} \\ &= \sqrt{8} \end{aligned}$$

$$\begin{aligned} d_E(T_1, B) &= \sqrt{(5-2)^2 + (0-0)^2 + (5-3)^2 + (2-1)^2} \\ &= \sqrt{9+0+4+1} \\ &= \sqrt{14} \end{aligned}$$

$$d_E(T_1, C) = \sqrt{35}$$

$$d_E(T_1, D) = \sqrt{49}$$

$\Rightarrow$  Classify  $T_1$  as Fruit.

For  $T_2$ :

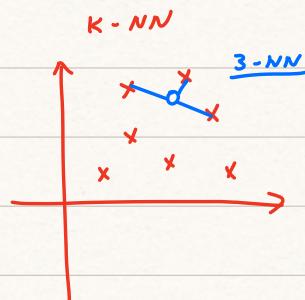
$$d_E(T_2, A) = \sqrt{14}$$

$$d_E(T_2, B) = \sqrt{40}$$

$$d_E(T_2, C) = \sqrt{11}$$

$$d_E(T_2, D) = \sqrt{9}$$

$\Rightarrow$  Classify  $T_2$  as computer



- (ii). Using the **Manhattan distance** measure, classify the test instances using the **3-NN** method, for the three weightings we discussed in the lectures: *majority class, inverse distance, inverse linear distance.*

$$d_M(A, B) = \sum_k |a_k - b_k|$$

$$d_M(T_1, A) = |4-2| + |0-0| + |1-3| + |1-1| \\ = 2+0+2+0$$

$$= 4$$

$$d_M(T_1, B) = 6$$

$$d_M(T_1, C) = 9$$

$$d_M(T_1, D) = 11$$

*Classification: (3 weightings)*

① Majority Class (equal weights)

2 Fruit, 1 Computer

⇒ Classify as Fruit

② Inverse distance ( $w = \frac{1}{d+1}$ ) ← avoid  $\frac{1}{0}$

Let  $\varepsilon = 1$

$$\text{for } A(\text{fruit}) : \frac{1}{4+1} = 0.2 \quad \left. \begin{array}{l} \text{"Score" for fruit} = 0.2 + 0.14 = 0.34 \\ \hline \end{array} \right\}$$

$$\text{for } B(\text{fruit}) : \frac{1}{6+1} = 0.14$$

$$\text{for } C(\text{comp}) : \frac{1}{9+1} = 0.1$$

0.34 (fruit) > 0.1 (comp) ⇒ Classify as fruit

③ Inverse linear distance ( $w_j = \frac{d_3 - d_j}{d_3 - d_1}$ ) (rescaling dist)  $\Rightarrow w_1 = \frac{d_3 - d_1}{d_3 - d_1} = 1$   
↑ nearest (in NN)      ↑ furthest       $w_3 = \frac{d_1 - d_3}{d_3 - d_1} = 0$   
 $\Rightarrow w_j \in [0, 1]$

**3-NN**

$$\left\{ \begin{array}{l} d_M(T_1, A) = 4 \\ d_M(T_1, B) = 6 \\ d_M(T_1, C) = 9 \end{array} \right.$$

For A (fruit) :  $\frac{9-4}{9-4} = 1$   
                           For B (fruit) :  $\frac{9-6}{9-4} = \frac{3}{5} = 0.6$   
                           For C (comp) :  $\frac{9-9}{9-4} = 0$

~~$d_M(T_1, D) = 11$~~

$$1 + 0.6 = 1.6$$

$1.6$  (fruit)  $> 0$  (comp)  $\Rightarrow$  Classify as fruit.

(iii). Can we do weighted k-NN using cosine similarity?

Yes, easier than distance, use cos similarity as weights directly.

<sup>(score)</sup>  
weight : cosine similarities.

All predictions:

Inst	Measure	k	Weight	Prediction
$T_1$	$d_E$	1	-	FRUIT
		3	Maj	FRUIT
		3	ID	FRUIT
		3	ILD	FRUIT
	$d_M$	1	-	FRUIT
		3	Maj	FRUIT
		3	ID	FRUIT
		3	ILD	FRUIT
	cos	1	-	FRUIT
		3	Maj	FRUIT
		3	Sum	FRUIT
$T_2$	$d_E$	1	-	COMPUTER
		3	Maj	FRUIT
		3	ID	FRUIT
		3	ILD	COMPUTER
	$d_M$	1	-	COMPUTER
		3	Maj	COMPUTER
		3	ID	COMPUTER
		3	ILD	COMPUTER
	cos	1	-	COMPUTER
		3	Maj	FRUIT
		3	Sum	FRUIT

2. Approximately 1% of women aged between 40 and 50 have breast cancer. 80% of mammogram screening tests detect breast cancer when it is there. 90% of mammograms DO NOT show breast cancer when it is NOT there<sup>1</sup>. Based on this information, complete the following table.

Cancer	Probability
No	99%
Yes	1%

Cancer	Test	Probability
Yes	Positive	80%
Yes	Negative	?
No	Positive	?
No	Negative	90%

$$\rightarrow P(P|C) = 0.8 \quad TP$$

$$\left\{ \begin{array}{l} P(N|C) = 1 - P(P|C) = 0.2 \quad FN \\ P(P|NC) = 1 - P(N|NC) = 0.1 \quad FP \\ P(N|NC) = 0.9 \quad TN \end{array} \right.$$

3. Based on the results in question 2, calculate the **marginal probability** of 'positive' results in a Mammogram Screening Test.

$$P(P) = ?$$

$\Rightarrow$  Law of total probability

$$\begin{aligned} P(P) &= \sum_{i \in \{C, NC\}} P(P|i) P(i) = \sum_{i \in \{C, NC\}} P(P,i) \\ &= P(P|C)P(C) + P(P|NC)P(NC) \\ &= 0.8 \times 0.01 + 0.1 \times 0.99 \\ &= 0.008 + 0.099 \\ &= 0.107 \end{aligned}$$

4. Based on the results in question 2, calculate  $P(\text{Cancer} = \text{'Yes'} | \text{Test} = \text{'Positive'})$ , using the Bayes Rule.

$$\begin{aligned} P(c|p) &= \frac{P(p|c)P(c)}{P(p)} \\ &= \frac{0.8 \times 0.01}{0.107} \\ &= 0.075 = 7.5\% \end{aligned}$$