CS 5990 (Advanced Data Mining) - Assignment #3
Maximum Points: 100 pts.

Bronco ID: 017314921
Last Name: Panda
First Name: Subham

Answers:

1.
Building the Euclidean distance table between samples:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$Sl.No.$	X	Y	$Class$	dis_1	dis_2	dis_3	dis_4	dis_5	dis_6
1	1	3	−	0	1	1.41	2	4.47	4.47
2	1	4	−	1	0	1	2.23	5	4.12
3	2	4	+	1.41	1	0	1.41	4.24	3.16
4	3	3	+	2	2.23	1.41	0	2.82	2.82
5	5	1	−	4.47	5	4.24	2.82	0	4
6	5	5	−	4.47	4.12	3.16	2.82	4	0

a)
Now, building the (LOO-CV) table for 1NN

$Sl.No.$	X	Y	$Class$	$1NN$	$Neighbors$
1	1	3	−	−	(1,4)
2	1	4	−	+	(1,3)
3	2	4	+	−	(1,4)
4	3	3	+	+	(2,4)
5	5	1	−	+	(3,3)
6	5	5	−	+	(3,3)

The Error Rate is $\frac{4}{6} = 0.667$

b)
Now, building the (LOO-CV) table for 3NN

$Sl. No.$	X	Y	$Class$	$3NN$	$Neighbors$
1	1	3	$-$	$+$	$[(1,3), (2,4), (3,3)]$
2	1	4	$-$	$+$	$[(1,3), (2,4), (3,3)]$
3	2	4	$+$	$-$	$[(1,4), (1,3), (3,3)]$
4	3	3	$+$	$-$	$[(2,4), (1,4), (1,3)]$
5	5	1	$-$	$+$	$[(3,3), (5,5), (2,4)]$
6	5	5	$-$	$+$	$[(3,3), (2,4), (5,1)]$

The Error Rate is $\frac{6}{6} = 1$

c)
Distance weights:

For P_1:
Nearest 3 neighbors-

$$\frac{1}{1^2} = \mathbf{1}\ (-)$$
$$\frac{1}{1.41^2} + \frac{1}{2^2} = 0.752\ (+)$$

Therefore, this will be $(+)$

For P_2:
Nearest 3 neighbors-

$$\frac{1}{1^2} = \mathbf{1}\ (-)$$
$$\frac{1}{1^2} + \frac{1}{2.23^2} = 1.201\ (+)$$

Therefore, this will be $(+)$

For P_3:
Nearest 3 neighbors-

$$\frac{1}{1^2} + \frac{1}{1.41^2} = \mathbf{1.5}\ (-)$$
$$\frac{1}{1.41^2} = 0.50\ (+)$$

Therefore, this will be $(-)$

For P_4:
Nearest 3 neighbors-

$$\frac{1}{2^2} + \frac{1}{2.23^2} = 0.451\ (-)$$
$$\frac{1}{1.41^2} = \mathbf{0.50}\ (+)$$

Therefore, this will be $(-)$

For P_5:
Nearest 3 neighbors-

$$\frac{1}{2.82^2} + \frac{1}{4.24^2} = \mathbf{0.181}\ (+)$$
$$\frac{1}{4^2} = 0.0625\ (-)$$

Therefore, this will be $(+)$

For P_6:
Nearest 3 neighbors-

$$\frac{1}{2.82^2}+\frac{1}{3.16^2}=\mathbf{0.225}\ (+)$$
$$\frac{1}{4^2}=0.0625\ (-)$$

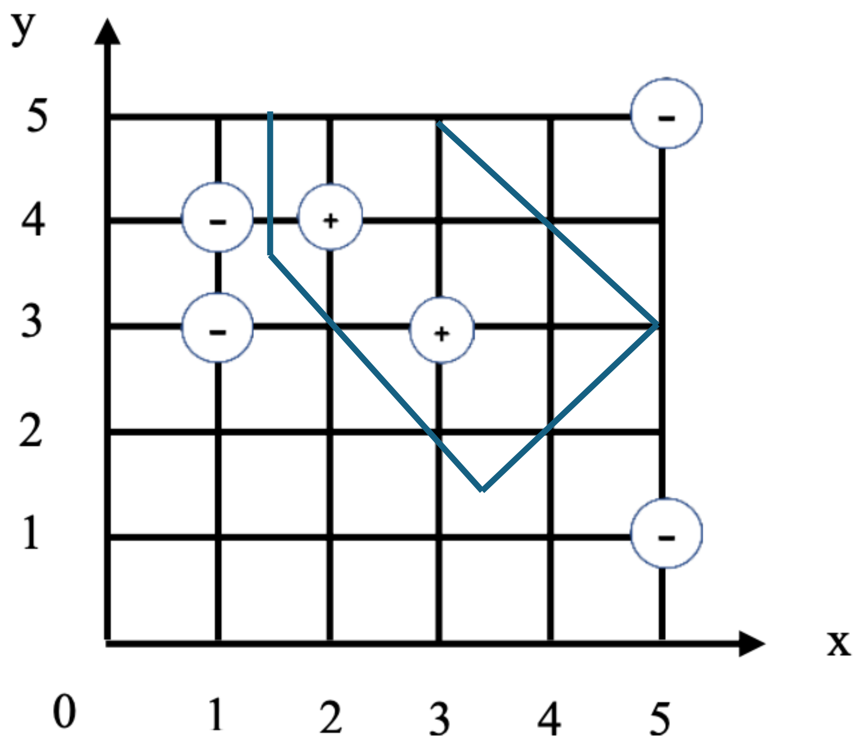
Therefore, this will be $(+)$

Now, building the (LOO-CV) table for 3NN(DW)

$Sl.No.$	X	Y	$Class$	$3NN(DW)$	$Neighbors$
1	1	3	$-$	$-$	$[(1,3),(2,4),(3,3)]$
2	1	4	$-$	$+$	$[(1,3),(2,4),(3,3)]$
3	2	4	$+$	$-$	$[(1,4),(1,3),(3,3)]$
4	3	3	$+$	$+$	$[(2,4),(1,4),(1,3)]$
5	5	1	$-$	$+$	$[(3,3),(5,5),(2,4)]$
6	5	5	$-$	$+$	$[(3,3),(2,4),(5,1)]$

The Error Rate is $\mathbf{\frac{4}{6}=0.667}$

d)
The decision boundary learned by 1NN algorithm (pseudo-model) is



2.
Using KNN with $K = 3$ to classify test sample $(t1 = 1, t2 = 2, t3 = 3, t4 = 4, t5 = 4)$ with $L1$ norm for distance computations:

Let, $t = (t1 = 1, t2 = 2, t3 = 3, t4 = 4, t5 = 4)$

L1 norm distances between test sample and training dataset would be:

$$P_{t1} = |1-2| + |2-3| + |3-4| + |4-5| + |4-5| = 5$$

$$P_{t2} = |1-0| + |2-1| + |3-2| + |4-3| + |4-5| = 5$$

$$P_{t3} = |1-2| + |2-2| + |3-2| + |4-2| + |4-4| = 4$$

$$P_{t4} = |1-0| + |2-1| + |3-2| + |4-3| + |4-5| = 5$$

$$P_{t5} = |1-4| + |2-2| + |3-4| + |4-4| + |4-4| = 4$$

Therefore, the **test sample for $\boldsymbol{K = 3}$ would be classified as $\boldsymbol{Sell}$** with the 3 nearest neighbors being $P_3 = Sell, P_4 = Sell, P_5 = Sell$

3.

KNN Program GitHub Link

4.

Naïve Bayes Approach

$$P(+ \mid A = 0, B = 1, C = 0) = P(A = 0, B = 1, C = 0 \mid +) * P(+)$$

$$\therefore P(A = 0 \mid +) * P(B = 1 \mid +) * P(C = 0 \mid +) * P(+)$$

$$= (2/5) * (1/5) * (1/5) * (5/10) = 0.008$$

$$P(- \mid A = 0, B = 1, C = 0) = P(A = 0, B = 1, C = 0 \mid -) * P(-)$$

$$\therefore P(A = 0 \mid -) * P(B = 1 \mid -) * P(C = 0 \mid -) * P(-)$$

$$= (3/5) * (2/5) * (0/5) * (5/10) = 0$$

Therefore, $(A = 0, B = 1, C = 0)$ would be classified as $+$.

Now, estimating the probabilities using m-estimate with $p = \frac{1}{2}$ and $m = 4$

$$\textbf{m−estimate:} P(A_i|C) = \frac{N_{ic} + mp}{N_c + m}$$

$$P(+|A = 0) = \frac{2+2}{5+4} = \frac{4}{9}$$

$$P(+|B = 1) = \frac{1+2}{5+4} = \frac{3}{9}$$

$$P(+|C = 0) = \frac{1+2}{5+4} = \frac{3}{9}$$

$$P(-|A = 0) = \frac{3+2}{5+4} = \frac{5}{9}$$

$$P(-|B = 1) = \frac{2+2}{5+4} = \frac{4}{9}$$

$$P(-|C = 0) = \frac{0+2}{5+4} = \frac{2}{9}$$

Now using the m-estimate probabilities calculating the naïve bayes classifier for the test sample:

$$P(+ \mid A = 0, B = 1, C = 0)$$

$$= (4/9) * (3/9) * (3/9) * (5/10) = 0.0247$$

$$P(- \mid A = 0, B = 1, C = 0)$$

$$= (5/9) * (4/9) * (2/9) * (5/10) = 0.0274$$

Now normalizing the scores:

For $+$,

$$\frac{0.0247}{0.0247 + 0.0274} = 0.474$$

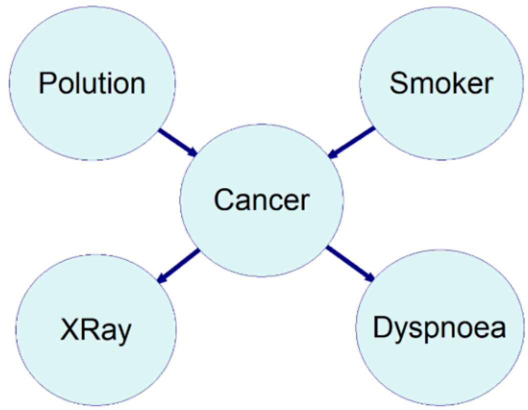
For $+$,

$$\frac{0.0274}{0.0247 + 0.0274} = \mathbf{0.525}$$

After using the m-estimate values for naïve bayes classification for the sample test the most probable classification is –

Therefore, $(A = 0, B = 1, C = 0)$ would be classified as $-$.

5.



a) The conditional probability tables for each variable are:

$Pollution$
$L\ =\ 11/20\ =\ 0.55$
$H\ =\ 9/20\ =\ 0.45$

$Smoker$
$False\ =\ 13/20\ =\ 0.65$
$True\ =\ 7/20\ =\ 0.35$

$Cancer$				
	$P = L$	$P = L$	$P = H$	$P = H$
	$S = False$	$S = True$	$S = False$	$S = True$
$False$	$7/8\ =\ 0.875$	$2/3\ =\ 0.667$	$3/5\ =\ 0.6$	$2/4\ =\ 0.5$
$True$	$1/8\ =\ 0.125$	$1/3\ =\ 0.333$	$2/5\ =\ 0.4$	$2/4\ =\ 0.5$

$Xray$		
	$C\ =\ False$	$C\ =\ True$
Neg	$10/14\ =\ 0.714$	$2/6\ =\ 0.333$
Pos	$4/14\ =\ 0.285$	$4/6\ =\ 0.666$

$Dyspnoea$		
	$C = False$	$C = True$
$False$	$10/14\ =\ 0.714$	$2/6\ =\ 0.333$
$True$	$4/14\ =\ 0.285$	$4/6\ =\ 0.666$

b)
Given,

$(\mathrm{P} = \mathrm{L}, \mathrm{S} = \mathrm{True}, \mathrm{C} = \mathrm{True}, \mathrm{X} = \mathrm{pos})$

$P(P,S,C,X,D) = P(P) * P(S) * P(C|P,S) * P(X|C) * P(D|C)$

For:

$$\Rightarrow P(P = L) * P\ (S = True) * P(C = True|P = L, S = True) * P(X = Pos|C = True) * P(D|\ C = True)$$

Now for $D = False$:

$\Rightarrow 0.55 * 0.35 * 0.333 * 0.666 * P(D = False|C = True)$

$\Rightarrow\ 0.0426 * 0.333 = 0.014$

Now for $D = True$:

$\Rightarrow 0.55 * 0.35 * 0.333 * 0.666 * P(D = True|C = True)$

$\Rightarrow\ 0.0426 * 0.666 = 0.0283$

Now Normalizing,

For False:

$$\frac{0.014}{0.014 + 0.0283} = 0.3309$$

For True:

$$\boldsymbol{\frac{0.0283}{0.014 + 0.0283} = 0.6690}$$

Therefore, the patient is likely to have Dyspnoea.

6.

[Naive Bayes Program GitHub Link]