Kumar Ayush
16MT30009

Assignment 1

Problem 1

	$X_1 = 1$	$X_1 = 2$	#
$\#Y = 1$	31	22	53
$\#Y = 2$	40	19	59
$\#Y = 3$	34	54	88
#	105	95	

Calculating probabilities

$P(Y = 1|X_1 = 1) = \dfrac{31}{105}$, $P(Y = 1|X_1 = 2) = \dfrac{22}{95}$, $P(Y = 2|X_1 = 2) = \dfrac{40}{105}$,

$P(Y = 2|X_1 = 2) = \dfrac{19}{95}$, $P(Y = 3|X_1 = 1) = \dfrac{34}{105}$, $P(Y = 3|X_1 = 2) = \dfrac{54}{95}$

	$X_2 = 1$	$X_2 = 2$	#
$\#Y = 1$	25	28	53
$\#Y = 2$	26	33	59
$\#Y = 3$	45	43	88
#	96	104	

$P(Y = 1|X_2 = 1) = \dfrac{25}{96}$, $P(Y = 1|X_2 = 2) = \dfrac{28}{104}$, $P(Y = 2|X_2 = 2) = \dfrac{36}{96}$,

$P(Y = 2|X_2 = 2) = \dfrac{33}{104}$, $P(Y = 3|X_2 = 1) = \dfrac{45}{96}$, $P(Y = 3|X_2 = 2) = \dfrac{43}{104}$

	$X_3 = A$	$X_3 = B$	$X_3 = C$	$X_3 = D$	#
$\#Y = 1$	35	15	1	2	53
$\#Y = 2$	14	25	15	5	59
$\#Y = 3$	1	10	34	43	88
#	50	50	50	50	

$P(Y = 1|X_3 = A) = \dfrac{35}{50}$, $P(Y = 1|X_3 = B) = \dfrac{15}{50}$, $P(Y = 2|X_3 = C) = \dfrac{1}{50}$,

$P(Y = 2|X_3 = D) = \dfrac{2}{50}$, $P(Y = 3|X_3 = A) = \dfrac{14}{50}$, $P(Y = 3|X_3 = B) = \dfrac{25}{50}$

$P(Y=1|X_3=C)=\frac{15}{50}$, $P(Y=1|X_3=D)=\frac{5}{50}$, $P(Y=2|X_3=A)=\frac{1}{50}$,
$P(Y=2|X_3=B)=\frac{10}{50}$, $P(Y=3|X_3=C)=\frac{34}{50}$, $P(Y=3|X_3=D)=\frac{43}{104}$

Entropy:

$$H=\sum_{i=1}^{n} P(X_i)\log_2(P(X_i))$$

$$H_{initial}=\frac{53}{200}\log_2\frac{53}{200}+\frac{59}{200}\log_2\frac{59}{200}+\frac{88}{200}\log_2\frac{88}{200}=1.5484$$
$$H_{X_1=1}=\frac{31}{105}\log_2\frac{31}{105}+\frac{40}{105}\log_2\frac{40}{105}+\frac{34}{105}\log_2\frac{34}{105}=1.5768$$
$$H_{X_1=2}=\frac{22}{95}\log_2\frac{22}{95}+\frac{19}{95}\log_2\frac{19}{95}+\frac{54}{95}\log_2\frac{54}{95}=1.4163$$
$$H_{X_2=1}=\frac{25}{96}\log_2\frac{25}{96}+\frac{26}{96}\log_2\frac{26}{96}+\frac{45}{96}\log_2\frac{45}{96}=1.5282$$
$$H_{X_2=2}=\frac{28}{104}\log_2\frac{28}{104}+\frac{33}{104}\log_2\frac{33}{104}+\frac{43}{104}\log_2\frac{43}{104}=1.56197$$
$$H_{X_3=A}=\frac{35}{50}\log_2\frac{35}{50}+\frac{14}{50}\log_2\frac{14}{50}+\frac{1}{50}\log_2\frac{1}{50}=0.98729$$
$$H_{X_3=B}=\frac{15}{50}\log_2\frac{15}{50}+\frac{25}{50}\log_2\frac{25}{50}+\frac{10}{50}\log_2\frac{10}{50}=1.4854$$
$$H_{X_3=C}=\frac{1}{50}\log_2\frac{1}{50}+\frac{15}{50}\log_2\frac{15}{50}+\frac{34}{50}\log_2\frac{34}{50}=1.0123$$
$$H_{X_3=D}=\frac{2}{50}\log_2\frac{5}{50}+\frac{5}{50}\log_2\frac{5}{50}+\frac{40}{50}\log_2\frac{43}{50}=0.7050$$
$$H_{X_3=\sim A}=\frac{18}{150}\log_2\frac{18}{150}+\frac{45}{150}\log_2\frac{45}{150}+\frac{87}{150}\log_2\frac{87}{150}=1.3439$$
$$H_{X_3=\sim B}=\frac{38}{50}\log_2\frac{38}{150}+\frac{34}{150}\log_2\frac{34}{150}+\frac{78}{150}\log_2\frac{78}{150}=1.4777$$
$$H_{X_3=\sim C}=\frac{52}{150}\log_2\frac{52}{150}+\frac{44}{150}\log_2\frac{44}{150}+\frac{54}{150}\log_2\frac{54}{150}=1.5794$$
$$H_{X_3=\sim D}=\frac{51}{150}\log_2\frac{51}{150}+\frac{54}{150}\log_2\frac{54}{150}+\frac{45}{150}\log_2\frac{45}{150}=1.58087$$
$$H_{X_3=(A,B)}=\frac{50}{150}\log_2\frac{50}{150}+\frac{39}{150}\log_2\frac{39}{150}+\frac{11}{150}\log_2\frac{11}{150}=1.3800$$
$$H_{X_3=(A,C)}=\frac{36}{100}\log_2\frac{36}{100}+\frac{29}{100}\log_2\frac{29}{100}+\frac{35}{100}\log_2\frac{35}{100}=1.5786$$

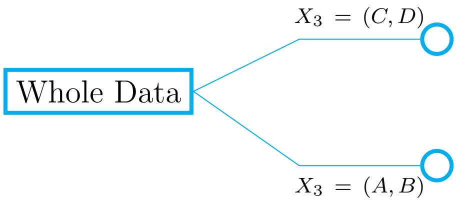
$$H_{X_3=(A,D)} = \frac{37}{100}\log_2\frac{37}{100} + \frac{19}{100}\log_2\frac{19}{100} + \frac{44}{100}\log_2\frac{44}{100} = 1.5071$$
$$H_{X_3=(B,C)} = \frac{16}{100}\log_2\frac{16}{100} + \frac{40}{100}\log_2\frac{40}{100} + \frac{44}{100}\log_2\frac{44}{100} = 1.4729$$
$$H_{X_3=(B,D)} = \frac{17}{100}\log_2\frac{17}{100} + \frac{30}{100}\log_2\frac{30}{100} + \frac{53}{100}\log_2\frac{53}{100} = 1.4411$$
$$H_{X_3=(C,D)} = \frac{3}{100}\log_2\frac{3}{100} + \frac{20}{100}\log_2\frac{20}{100} + \frac{77}{100}\log_2\frac{77}{100} = 0.9064$$

Information Gain

$$IG = H_{initial} - (P_{split_1} * H_{split_1} + P_{split_2} * H_{split_2})$$

$$IG_{X_1} = 1.5484 - \left(\frac{105}{200} * 1.5768 + \frac{95}{200} * 1.4163\right) = 0.29365$$
$$IG_{X_2} = 1.5484 - \left(\frac{96}{200} * 1.5282 + \frac{104}{200} * 1.56197\right) = 0$$
$$IG_{X_3=(A,\sim A)} = 1.5484 - \left(\frac{50}{200} * 0.98729 + \frac{150}{200} * 1.3439\right) = 0.2936$$
$$IG_{X_3=(B,\sim B)} = 1.5484 - \left(\frac{50}{200} * 1.4854 + \frac{150}{200} * 1.4777\right) = 0.0687$$
$$IG_{X_3=(C,\sim C)} = 1.5484 - \left(\frac{50}{200} * 1.0123 + \frac{150}{200} * 1.5794\right) = 0.1107$$
$$IG_{X_3=(D,\sim D)} = 1.5484 - \left(\frac{50}{200} * 0.7050 + \frac{150}{200} * 1.58087\right) = 0.1864$$
$$IG_{X_3=((A,B),(C,D))} = 1.5484 - \left(\frac{100}{200} * 1.3800 + \frac{100}{200} * 0.9064\right) = 0.4052$$
$$IG_{X_3=((A,C),(B,D))} = 1.5484 - \left(\frac{100}{200} * 1.5786 + \frac{100}{200} * 1.4411\right) = 0.3854$$
$$IG_{X_3=((A,D),(B,C))} = 1.5484 - \left(\frac{100}{200} * 1.5071 + \frac{100}{200} * 1.4729\right) = 0.06159$$

Maximum information gain happens when we split X_3 = ((A, B), (C, D))

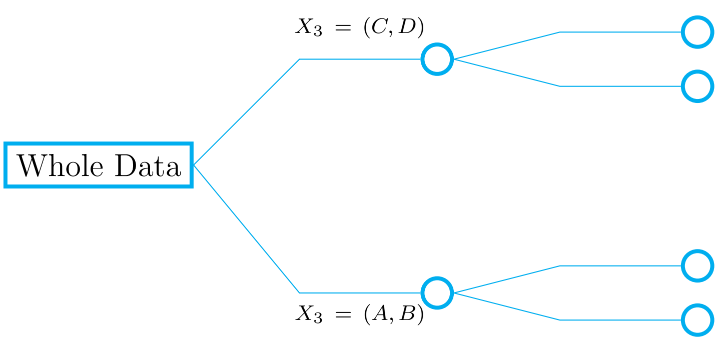


$$P(Y=1|X_1=1) = \frac{31}{105},\ P(Y=1|X_1=2) = \frac{22}{95},\ P(Y=2|X_1=2) = \frac{40}{105},$$

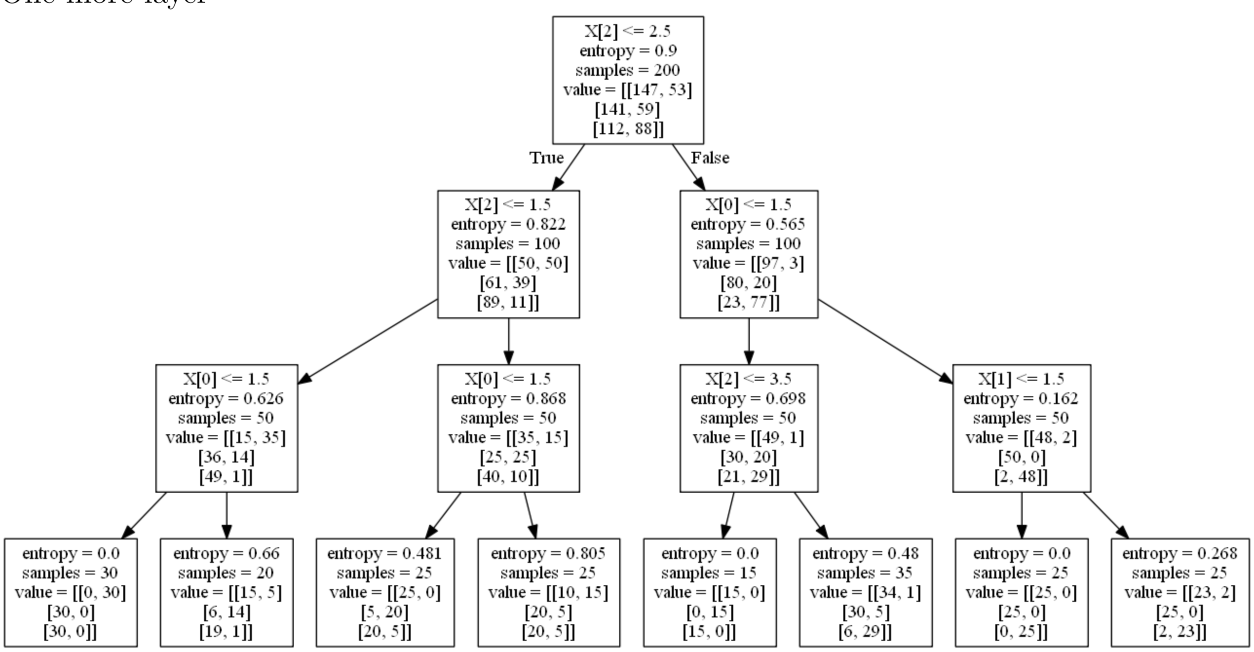
	$X_1 = 1$	$X_1 = 2$	#
$\#Y = 1$	30	20	50
$\#Y = 2$	20	19	39
$\#Y = 3$	5	6	11
#	55	45	

$$P(Y = 2|X_1 = 2) = \frac{19}{95},\ P(Y = 3|X_1 = 1) = \frac{34}{105},\ P(Y = 3|X_1 = 2) = \frac{54}{95}$$

Similarly doing these calculations till depth = 3 i.e. 2 more iterations



One more layer



Accuracy = 0.855

Problem 2

Case 1: X = (2, 1, A)

$$P(X_1 = 2 \mid Y = 1) = \frac{P(X_1 = 2 \cap Y = 1)}{P(Y = 1)} = \frac{\frac{12}{170}}{\frac{43}{170}} = 0.28$$

$$P(X_2 = 1 \mid Y = 1) = \frac{P(X_2 = 1 \cap Y = 1)}{P(Y = 1)} = \frac{\frac{15}{170}}{\frac{43}{170}} = 0.35$$

$$P(X_3 = A \mid Y = 1) = \frac{P(X_3 = A \cap Y = 1)}{P(Y = 1)} = \frac{\frac{32}{170}}{\frac{43}{170}} = 0.74$$

$$\implies P(X|Y = 1) = 0.072$$

$$P(Y = 1 \mid X) = \frac{P(Y = 1 \cap X)}{P(Y = 1)} = 0.45$$

$$P(Y = 2 \mid X) = \frac{P(Y = 2 \cap X)}{P(Y = 1)} = 0.086$$

$$P(Y = 3 \mid X) = \frac{P(Y = 3 \cap X)}{P(Y = 1)} = 0.04$$

Confidence = 0.45 Final Prediction (Max P(Y = i — X)) $\implies Y = 1$

Case 2:$X = (2, 1, B)$P(X — Y = 1) = $\dfrac{P(X \cap Y = 1)}{P(Y = 1)} = \dfrac{12 * 8 * 15}{43 * 43 * 43} = 0.018$

$$P(Y = 1|X) = \frac{P(Y = 1 \cap X)}{P(X)} = 0.12$$

$$P(X|Y = 2) = \frac{P(X \cap Y = 2)}{P(Y = 2)} = \frac{11 * 22 * 12}{45 * 45 * 45} = 0.031$$

$$P(Y = 2|X) = \frac{P(Y = 2 \cap X)}{P(X)} = 0.023$$

$$P(X|Y = 3) = \frac{P(X \cap Y = 3)}{P(Y = 3)} = \frac{48 * 39 * 9}{82 * 82 * 82} = 0.031$$

$$P(Y = 3|X) = \frac{P(Y = 3 \cap X)}{P(X)} = 0.39$$

Confidence = 0.39 Final Prediction (Max P(Y = i — X)) $\implies Y = 3$

Case 3:$X = (2, 1, D)$P(X — Y = 1) = $\dfrac{P(X \cap Y = 1)}{P(Y = 1)} = \dfrac{12 * 2 * 15}{43 * 43 * 43} = 0.0045$

$P(Y = 1|X) = \dfrac{P(Y = 1 \cap X)}{P(X)} = 0.026$

$P(X|Y = 2) = \dfrac{P(X \cap Y = 2)}{P(Y = 2)} = \dfrac{11 * 5 * 12}{45 * 45 * 45} = 0.0072$

$P(Y = 2|X) = \dfrac{P(Y = 2 \cap X)}{P(X)} = 0.044$

$P(X|Y = 3) = \dfrac{P(X \cap Y = 3)}{P(Y = 3)} = \dfrac{48 * 39 * 38}{82 * 82 * 82} = 0.0129$

$P(Y = 3|X) = \dfrac{P(Y = 3 \cap X)}{P(X)} = 0.89$

Confidence = 0.89 Final Prediction (Max P(Y = i — X)) $\implies Y = 3$

Case 4 : $X = (1, 1, C)$P(X — Y = 1) = $\dfrac{P(X \cap Y = 1)}{P(Y = 1)} = \dfrac{31 * 1 * 15}{43 * 43 * 43} = 0.0058$

$P(Y = 1|X) = \dfrac{P(Y = 1 \cap X)}{P(X)} = 0.02$

$P(X|Y = 2) = \dfrac{P(X \cap Y = 2)}{P(Y = 2)} = \dfrac{34 * 12 * 12}{45 * 45 * 45} = 0.04$

$P(Y = 2|X) = \dfrac{P(Y = 2 \cap X)}{P(X)} = 0.18$

$P(X|Y = 3) = \dfrac{P(X \cap Y = 3)}{P(Y = 3)} = \dfrac{34 * 39 * 34}{82 * 82 * 82} = 0.082$

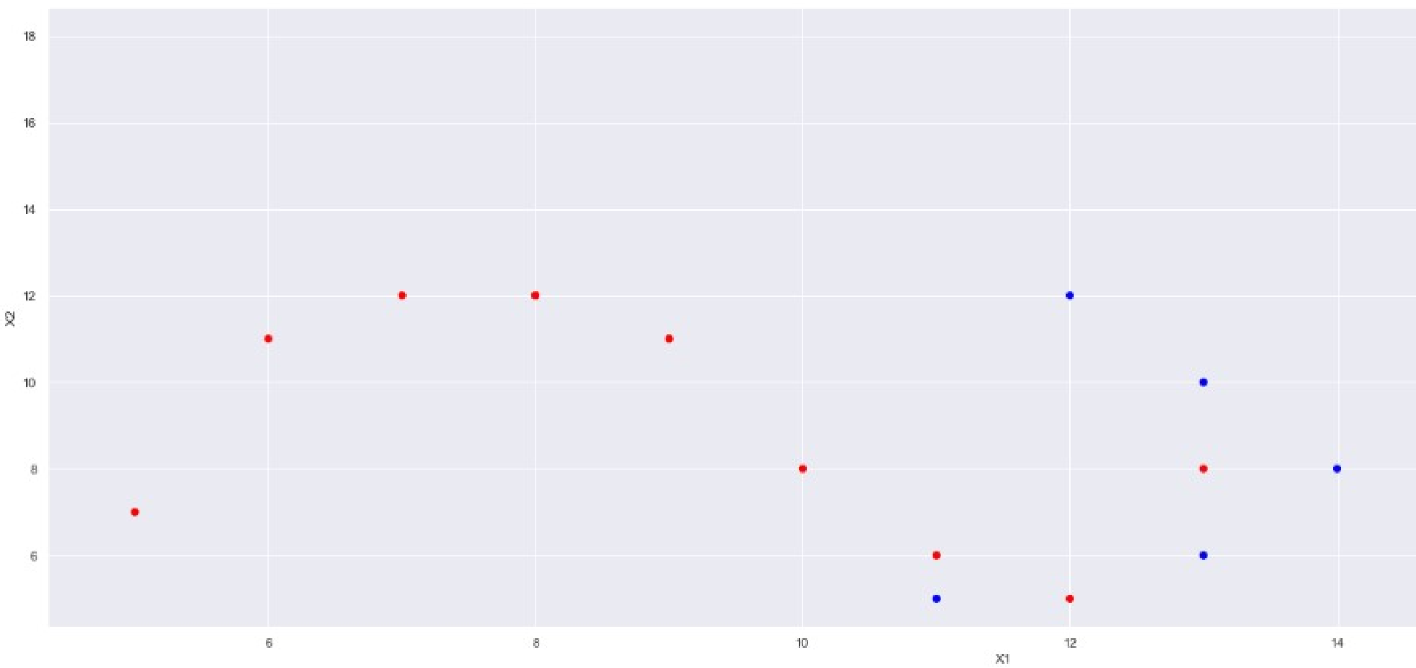
$P(Y = 3|X) = \dfrac{P(Y = 3 \cap X)}{P(X)} = 0.68$

Confidence = 0.68 Final Prediction (Max P(Y = i — X)) $\implies Y = 3$

Accuracy = 0.74

Problem 3

Plot



Looking at plot it is clear that a good split can be $X_1 = 10$
Split will be done only if information gain is positive

$H_{initial} = -\frac{10}{20}\log_2\frac{10}{20} - \frac{10}{20} * \log_2\frac{10}{20} = 1$

$H(X_1 \leq 11) = -\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9} * \log_2\frac{1}{9} = 0.5032$

$H(X_1 > 11) = -\frac{2}{11}\log_2\frac{2}{11} - \frac{9}{11} * \log_2\frac{9}{11} = 0.6840$

$\text{IG} = 1 - \left(\frac{9}{20} * 0.5032 - \frac{11}{20} * 0.6840\right) = 39736$

Best Split: $X_1 = 11$

Problem 4

$\mu_{X_1} = 11.65$
$\mu_{X_1} = 9.55$
$stdDev_{X_1} = 3.674593$
$stdDev_{X_2} = 3.379115$

Algorithm 1: Normal Distribution

```
import numpy as np
import pandas as pd
from sklearn import tree
from sklearn.preprocessing import OneHotEncoder
from sklearn.metrics import accuracy_score
from sklearn import tree
import pydot
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
%matplotlib inline

from IPython.display import Image

X = np.array([
    [5, 7],
    [7, 12],
    [12, 5],
    [10, 8],
    [6, 11],
    [13, 8],
    [8, 12],
    [9, 11],
    [11, 6],
    [8, 12],
    [13, 6],
    [14, 8],
    [17, 15],
    [15, 9],
    [13, 10],
    [11, 5],
    [16, 18],
    [15, 7],
    [12, 12],
    [18, 9]
])

Y = np.array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
X = pd.DataFrame(X,columns=['X1','X2'])
Y = pd.DataFrame(Y,columns=['Y1'])

from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
y_pred = gnb.fit(X, Y).predict(X)
# y_pred is [1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2]

proba_y_pred = gnb.predict_proba(X)
'''
array([[9.99859060e-01, 1.40939825e-04],
       [9.96573067e-01, 3.42693259e-03],
```

```
       [3.94899482e-01, 6.05100518e-01],
       [9.12468590e-01, 8.75314104e-02],
       [9.99325593e-01, 6.74406701e-04],
       [2.89854322e-01, 7.10145678e-01],
       [9.87238614e-01, 1.27613857e-02],
       [9.65393545e-01, 3.46064553e-02],
       [7.15417535e-01, 2.84582465e-01],
       [9.87238614e-01, 1.27613857e-02],
       [2.35537973e-01, 7.64462027e-01],
       [1.35388601e-01, 8.64611399e-01],
       [2.31921435e-03, 9.97680786e-01],
       [6.07414639e-02, 9.39258536e-01],
       [2.77679262e-01, 7.22320738e-01],
       [6.57758049e-01, 3.42241951e-01],
       [6.76505314e-04, 9.99323495e-01],
       [5.47271837e-02, 9.45272816e-01],
       [4.16302198e-01, 5.83697802e-01],
       [6.28435643e-03, 9.93715644e-01]])
'''
```

For $X_2 = 7$ when $X_1 = 5$ predicted probability falls lowest, hence X_2