w2

Graded

Group

Mihret Kemal
Feven Tefera
Tagore Kosireddy
...and 1 more
View or edit group

Total Points

59.5 / 74 pts

Question 1

Question 1 — **27.5** / 42 pts

1.1 **Q1(a)** — **7.5** / 18 pts

- ✓ **– 10.5 pts** 14 incorrect probabilities

Make you understand how to calculate these probabilities for the midterm exam.

1. OK
2. OK
3. OK
4. OK

1.2 **Q1(b)** — **20** / 24 pts

- ✓ **– 4 pts** Correct form of calculations, incorrect values (all test samples)

Question 2

Question 2 **32** / 32 pts

2.1 Q2(a)i **4** / 4 pts

- **− 0 pts** Correct

2.2 Q2(a)ii **5** / 5 pts

- **− 0 pts** Correct

2.3 Q2(a)iii **5** / 5 pts

- **− 0 pts** Correct

2.4 Q2(b) **2** / 2 pts

- **− 0 pts** Correct

2.5 Q2(c)i **4** / 4 pts

- **− 0 pts** Correct

2.6 Q2(c)ii **5** / 5 pts

- **− 0 pts** Correct

2.7 Q2(c)iii **5** / 5 pts

- **− 0 pts** Correct

2.8 Q2(d) **2** / 2 pts

- **− 0 pts** Correct

Question 3

General **0** / 0 pts

- **− 0 pts** No comments

No questions assigned to the following page.

Group Name	Charlie
Group Member 1	Feven Tefera
Group Member 2	Michael Ngala
Group Member 3	Mihret Kemel
Group Member 4	Tagore Kosireddy

Q1

(a) Estimate the conditional probabilities needed for Naïve Bayes classification using Laplace smoothing, where $\alpha = 1$ and β is size of variable's domain. Estimate an unsmoothed prior probability.

$$P(Y = y_k) = \frac{\#(Y = y_k)}{|domain(Y)|}$$

$$P(X_i = j \,|\, Y = y_k) = \frac{\#(X_i = j, Y = y_k) + 1}{\#(Y = y_k) + |domain(X_i)|}$$

Fill in the following tables (**report as fractions**):

Prior	Prob.
$P(apple)$	$\frac{14}{25}$
$P(orange)$	$\frac{11}{25}$

Cond.	Prob.
$P(Wt = 0 \,\|\, apple)$	$\frac{11}{28}$
$P(Wt = 1 \,\|\, apple)$	$\frac{1}{5}$
$P(Wt = 0 \,\|\, orange)$	$\frac{1}{5}$
$P(Wt = 1 \,\|\, orange)$	$\frac{4}{11}$

Cond.	Prob.
$P(Ht = 0 \,\|\, apple)$	$\frac{2}{5}$
$P(Ht = 1 \,\|\, apple)$	$\frac{1}{7}$
$P(Ht = 2 \,\|\, apple)$	$\frac{4}{21}$
$P(Ht = 0 \,\|\, orange)$	$\frac{3}{22}$
$P(Ht = 1 \,\|\, orange)$	$\frac{5}{17}$
$P(Ht = 2 \,\|\, orange)$	$\frac{5}{18}$

Cond.	Prob.
$P(Wid = 0 \,\|\, apple)$	$\frac{11}{31}$
$P(Wid = 1 \,\|\, apple)$	$\frac{3}{19}$
$P(Wid = 2 \,\|\, apple)$	$\frac{3}{17}$
$P(Wid = 0 \,\|\, orange)$	$\frac{2}{7}$
$P(Wid = 1 \,\|\, orange)$	$\frac{1}{4}$
$P(Wid = 2 \,\|\, orange)$	$\frac{1}{7}$

Question assigned to the following page: 1.2

(b) Report the predicted class on the test samples using the estimated parameters above.

Sample Num.	Test Data [Weight, Height, Width]	Prediction
1	[1, 0, 0]	apple
2	[0, 0, 1]	apple
3	[1, 2, 0]	orange
4	[0, 1, 1]	orange

<u>Show the calculations</u>, writing out the probabilities that go into making the prediction.

1.Given $Wt = 1\,, Ht = 0\,, Wid = 0$ What is the Prediction?
For this let's calculate
$P(apple \,|\, Wt = 1\,, Ht = 0\,, Wid = 0) = P(Wt = 1 \,|\, apple) * P(Ht = 0 \,|\, apple) * P(Wid = 0 \,|\, apple) * P(apple)$
$P(apple \,|\, Wt = 1\,, Ht = 0\,, Wid = 0) = \frac{1}{5} * \frac{2}{5} * \frac{11}{31} * \frac{14}{25} = 0.01589677419$
Now let's Calculate
$P(orange \,|\, Wt = 1\,, Ht = 0\,, Wid = 0) = P(Wt = 1 \,|\, orange) * P(Ht = 0 \,|\, orange) * P(Wid = 0 \,|\, orange) * P(orange)$
$P(orange \,|\, Wt = 1\,, Ht = 0\,, Wid = 0) = \frac{4}{11} * \frac{3}{22} * \frac{2}{7} * \frac{11}{25} = 0.00623376623$
Since $P(apple \,|\, Wt = 1\,, Ht = 0\,, Wid = 0) > P(orange \,|\, Wt = 1\,, Ht = 0\,, Wid = 0)$
The first prediction is apple.

2.Given $Wt = 0\,, Ht = 0\,, Wid = 1$ What is the Prediction?
For this let's calculate
$P(apple \,|\, Wt = 0\,, Ht = 0\,, Wid = 1) = P(Wt = 0 \,|\, apple) * P(Ht = 0 \,|\, apple) * P(Wid = 1 \,|\, apple) * P(apple)$
$P(apple \,|\, Wt = 0\,, Ht = 0\,, Wid = 1) = \frac{11}{28} * \frac{2}{5} * \frac{3}{19} * \frac{14}{25} = 0.01389473684$
Now let's Calculate
$P(orange \,|\, Wt = 0\,, Ht = 0\,, Wid = 1) = P(Wt = 0 \,|\, orange) * P(Ht = 0 \,|\, orange) * P(Wid = 1 \,|\, orange) * P(orange)$
$P(orange \,|\, Wt = 0\,, Ht = 0\,, Wid = 1) = \frac{1}{5} * \frac{3}{22} * \frac{1}{4} * \frac{11}{25} = 0.003$
Since $P(apple \,|\, Wt = 0\,, Ht = 0\,, Wid = 1) > P(orange \,|\, Wt = 0\,, Ht = 0\,, Wid = 1)$
The second prediction is apple.

3.Given $Wt = 1\,, Ht = 2\,, Wid = 0$ What is the Prediction?
For this let's calculate
$P(apple \,|\, Wt = 1\,, Ht = 2\,, Wid = 0) = P(Wt = 1 \,|\, apple) * P(Ht = 2 \,|\, apple) * P(Wid = 0 \,|\, apple) * P(apple)$
$P(apple \,|\, Wt = 1\,, Ht = 2\,, Wid = 0) = \frac{1}{5} * \frac{4}{21} * \frac{11}{31} * \frac{14}{25} = 0.00756989247$
Now let's Calculate
$P(orange \,|\, Wt = 1\,, Ht = 2\,, Wid = 0) = P(Wt = 1 \,|\, orange) * P(Ht = 2 \,|\, orange) * P(Wid = 0 \,|\, orange) * P(orange)$
$P(orange \,|\, Wt = 1\,, Ht = 2\,, Wid = 0) = \frac{4}{11} * \frac{5}{18} * \frac{2}{7} * \frac{11}{25} = 0.0126984127$

Question assigned to the following page: 1.2

Since $P(apple \,|\, Wt = 1\,, Ht = 2\,, Wid = 0) < P(orange \,|\, Wt = 1\,, Ht = 2\,, Wid = 0)$
The third prediction is orange.

4.Given $Wt = 0\,, Ht = 1\,, Wid = 1$ What is the Prediction?
For this let's calculate
$P(apple \,|\, Wt = 0\,, Ht = 1\,, Wid = 1) = P(Wt = 0 \,|\, apple) * P(Ht = 1 \,|\, apple) * P(Wid = 1 \,|\, apple) * P(apple)$
$P(apple \,|\, Wt = 0\,, Ht = 1\,, Wid = 1) = \frac{11}{28} * \frac{1}{7} * \frac{3}{19} * \frac{14}{25} = 0.00496240601$
Now let's Calculate
$P(orange \,|\, Wt = 0\,, Ht = 1\,, Wid = 1) = P(Wt = 0 \,|\, orange) * P(Ht = 1 \,|\, orange) * P(Wid = 1 \,|\, orange) * P(orange)$
$P(orange \,|\, Wt = 0\,, Ht = 1\,, Wid = 1) = \frac{1}{5} * \frac{5}{17} * \frac{1}{4} * \frac{11}{25} = 0.00647058823$
Since $P(apple \,|\, Wt = 0\,, Ht = 1\,, Wid = 1) < P(orange \,|\, Wt = 0\,, Ht = 1\,, Wid = 1)$
The fourth prediction is orange.

Question assigned to the following page: 2.1

Q2

(a) (14 points) Compute the information gain (based on entropy) for the two possible at- tributes. Show the form of the calculations, not just the final numbers.

i. Entropy before the split

Let's Calculate the individual probabilities of respective classes

$P(A) = \frac{2}{5}$, $P(B) = \frac{3}{10}$, $P(C) = \frac{3}{10}$

Entropy $= -\sum_{i=1}^{k} p_{m,i} \log_2 p_{m,i}$

Entropy $= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10}$

Entropy = 1.570950594

Question assigned to the following page: 2.2

ii. Information gain for variable, X_1

Information gain $= H(X_m) - (-\sum_{l=1}^{j} \frac{n_l}{n_m} H(X_{m,l}))$

Entropy for $N_{1,1} = -(\frac{43}{68}\log_2\frac{43}{68} + \frac{18}{68}\log_2\frac{18}{68} + \frac{7}{68}\log_2\frac{7}{68})$

Entropy for $N_{1,2} = -(\frac{37}{132}\log_2\frac{37}{132} + \frac{42}{132}\log_2\frac{42}{132} + \frac{53}{132}\log_2\frac{53}{132})$

Information gain $= 1.570950594 - (\frac{68}{200}(-(\frac{43}{68}\log_2\frac{43}{68} + \frac{18}{68}\log_2\frac{18}{68} + \frac{7}{68}\log_2\frac{7}{68})) + (\frac{132}{200}(-(\frac{37}{132}\log_2\frac{37}{132} + \frac{42}{132}\log_2\frac{42}{132} + \frac{53}{132}\log_2\frac{53}{132})))$

Information gain for variable, X_1 = 0.106145156

Question assigned to the following page: 2.3

iii. Information gain for variable, X_2

Information gain $= H(X_m) - (-\sum_{l=1}^{j} \frac{n_l}{n_m} H(X_{m,l}))$

Entropy for $N_{2,1} = -(\frac{46}{85}\log_2\frac{46}{85} + \frac{26}{85}\log_2\frac{26}{85} + \frac{13}{85}\log_2\frac{13}{85})$

Entropy for $N_{2,2} = -(\frac{14}{61}\log_2\frac{14}{61} + \frac{12}{61}\log_2\frac{12}{61} + \frac{35}{61}\log_2\frac{35}{61})$

Entropy for $N_{2,3} = -(\frac{20}{54}\log_2\frac{20}{54} + \frac{22}{54}\log_2\frac{22}{54} + \frac{12}{54}\log_2\frac{12}{54})$

Information gain $= 1.570950594 - (\frac{85}{200}(-(\frac{46}{85}\log_2\frac{46}{85} + \frac{26}{85}\log_2\frac{26}{85} + \frac{13}{85}\log_2\frac{13}{85})) + (\frac{61}{200}(-(\frac{14}{61}\log_2\frac{14}{61} + \frac{12}{61}\log_2\frac{12}{61} + \frac{35}{61}\log_2\frac{35}{61})) + (\frac{54}{200}(-(\frac{20}{54}\log_2\frac{20}{54} + \frac{22}{54}\log_2\frac{22}{54} + \frac{12}{54}\log_2\frac{12}{54})))$

Information gain for variable, X_2 = 0.123335088

Question assigned to the following page: 2.4

(b) (2 points) Which variable would be preferred to be included next in the decision tree?

We choose variable X_2 to be included in the decision tree because the information gain for the variable X_2 is greater than variable X_1.

Question assigned to the following page: 2.5

(c) (14 points) Compute the gain in GINI index for the two possible attributes. Show the form of the calculations, not just the final numbers.

i. GINI before the split

$GINI(X_m) = 1 - \sum_{j=1}^{k} p_{m,j}^2$

$GINI(X_m) = 1 - p(A)^2 - p(B)^2 - p(C)^2$

$GINI(X_m) = 1 - (\frac{2}{5})^2 - (\frac{3}{10})^2 - (\frac{3}{10})^2$

$GINI(X_m) \mathbf{= 0.66}$

Question assigned to the following page: 2.6

ii. gain in GINI for variable, X_1

$$gain\ in\ GINI(X_1) = GINI(X_m) - \sum_{l=1}^{j} \frac{n_l}{n_m} GINI(X_{m,l})$$

$GINI(N_{1,1}) = 1 - ((\frac{43}{68})^2 + (\frac{18}{68})^2 + (\frac{7}{68})^2)$

$GINI(N_{1,2}) = 1 - ((\frac{37}{132})^2 + (\frac{42}{132})^2 + (\frac{53}{132})^2)$

$gain\ in\ GINI(X_1) = 0.66 - ((\frac{68}{200})(1 - ((\frac{43}{68})^2 + (\frac{18}{68})^2 + (\frac{7}{68})^2)) + (\frac{132}{200})(1 - ((\frac{37}{132})^2 + (\frac{42}{132})^2 + (\frac{53}{132})^2)))$

$gain\ in\ GINI(X_1)$ **= 0.04845811**

Question assigned to the following page: 2.7

iii. gain in GINI for variable, X_2

$$gain\ in\ GINI(X_2) = GINI(X_m) - \sum_{l=1}^{j} \frac{n_l}{n_m} GINI(X_{m,l})$$

$GINI(N_{2,1}) = 1 - ((\frac{46}{85})^2 + (\frac{26}{85})^2 + (\frac{13}{85})^2)$

$GINI(N_{2,2}) = 1 - ((\frac{14}{61})^2 + (\frac{12}{61})^2 + (\frac{35}{61})^2)$

$GINI(N_{2,3}) = 1 - ((\frac{20}{54})^2 + (\frac{22}{54})^2 + (\frac{12}{54})^2)$

$gain\ in\ GINI(X_2) = 0.66 - ((\frac{85}{200})(1 - ((\frac{46}{85})^2 + (\frac{26}{85})^2 + (\frac{13}{85})^2)) + (\frac{61}{200})(1 - ((\frac{14}{61})^2 + (\frac{12}{61})^2 + (\frac{35}{61})^2)) + (\frac{54}{200})(1 - ((\frac{20}{54})^2 + (\frac{22}{54})^2 + (\frac{12}{54})^2)))$

$gain\ in\ GINI(X_2) = \mathbf{0.057640344}$

Question assigned to the following page: 2.8

(d) (2 points) Which variable would be preferred to include next in the decision tree?
We choose variable X_2 to include next in the decision tree because variable X_2 has greater gain in GINI INDEX than variable X_1.