National University of Computer and Emerging Sciences
Chiniot-Faisalabad Campus

# Data Mining (CS4038)

# Sessional-II Exam

Date: April 8th 2024

**Course Instructor(s)**

Ms. Ayesha Liaqat

**Total Time (Hrs):** 1

**Total Marks:** 60

**Total Questions:** 4

Solution

Roll No ____________ Section ____________ Student Signature ____________

**Vetted by:** NASREEN AKHTAR **Signature:**

**Attempt all the questions.**

***CLO # 2: Understand the nature of the data and apply data mining techniques to interpret the results.***

**Q1:** Given a dataset where we want to predict whether a customer will purchase a product based on their gender, age group, and preferred payment method. [10 marks]

| Gender | Age Group | Payment Method | Purchase |
|---|---|---|---|
| Male | Teen | Credit Card | No |
| Female | Adult | PayPal | Yes |
| Male | Adult | Cash | No |
| Female | Senior | Credit Card | Yes |
| Male | Teen | Cash | No |
| Female | Adult | PayPal | Yes |
| Female | Senior | Credit Card | Yes |
| Male | Teen | Cash | No |
| Male | Senior | PayPal | Yes |
| Female | Adult | Cash | No |

Demonstrate which attribute would you choose as the root node in a decision tree with multi-way splits using entropy impurity measure?

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} P_i(t) \log_2 P_i(t)$$

① Purchase

Ent(Purchase) $= -\frac{5}{10}\log_2(\frac{5}{10}) - \frac{5}{10}\log_2(\frac{5}{10})$

$= 0.5 + 0.5 = \boxed{1}$

| Purchase | |
|---|---|
| Yes | 5 |
| No | 5 |

② Ent(Gender)

+1 Ent(M/Gender) $= -\frac{1}{5}\log_2(\frac{1}{5}) - \frac{4}{5}\log_2(\frac{4}{5}) =$ +3

$= 0.464 + 0.257 = 0.721$

+1 Ent(F/Gender) $= 0.721$

+1 Ent(Gender) $= \frac{5}{10}(0.721) + \frac{5}{10}(0.721) = 0.721$

Gain info = Ent(Purchase) − Ent(Gender) = 1 − 0.721 = 0.279

| Gender | Male | Female |
|---|---|---|
| Yes | 1 | 4 |
| No | 4 | 1 |

③ Age Group

Ent(Teen/Agegroup) $= -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$ +3

Ent(Adult|Age group) = 1

Ent(Senior|Age group) = 0

Ent(Age Group) $= \frac{3}{10}(0) + \frac{4}{10}(1) + \frac{3}{10}(0) = 0.4$

Gain info = 1 − 0.4 = 0.6

| Age group | Teen | Adult | Senio |
|---|---|---|---|
| Yes | 0 | 2 | 3 |
| No | 3 | 2 | 0 |

④ Payment Method

+3

Ent(cc|PM) $= -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.385 + 0.533$

$= 0.9181$

Ent(Paypal|PM) = 0

Ent(Cash|PM) = 0

Ent(Payment Method) $= \frac{3}{10}(0.9181) + \frac{3}{10}(0) + \frac{4}{10}(0) = 0.2754$

Gain info $= 1 - 0.2754 = \boxed{0.7246}$

| Payment Method | cc | PayPal | Cash |
|---|---|---|---|
| Yes | 2 | 3 | 0 |
| No | 1 | 0 | 4 |

+1

So, As per the max Entropy and Gain info, Payment Method serve as the Root node for decision tree Construction

***CLO # 2: Understand the nature of the data and apply data mining techniques to interpret the results.***

**Q2:** Consider the following dataset, [10+5+10=25 marks]

a) Construct a kdtree with y-x split order using median method. Store datapoints at leaf nodes only and attributes at non-leaf nodes. Keep splitting if there are more than 2 points in any region.

| Points | X | Y | Class |
|---|---|---|---|
| P1 | 8 | 17 | - |
| P2 | 6 | 20 | + |
| P3 | 4 | 21 | - |
| P4 | 9 | 13 | + |
| P5 | 12 | 12 | + |
| P6 | 4 | 26 | + |
| P7 | 3 | 22 | - |
| P8 | 7 | 23 | - |

(i)

Sorted on y

| | |
|---|---|
| 12 | 12 |
| 9 | 13 |
| 8 | 17 |
| 6 | 20 |
| 4 | 21 |
| 3 | 22 |
| 7 | 23 |
| 4 | 26 |

Median = 20.5 +2

⇒ (ii) ≤ Sorted on x

| | |
|---|---|
| 6 | 20 |
| 8 | 17 |
| 9 | 13 |
| 12 | 12 |

M=8.5

+2

> Sorted on x

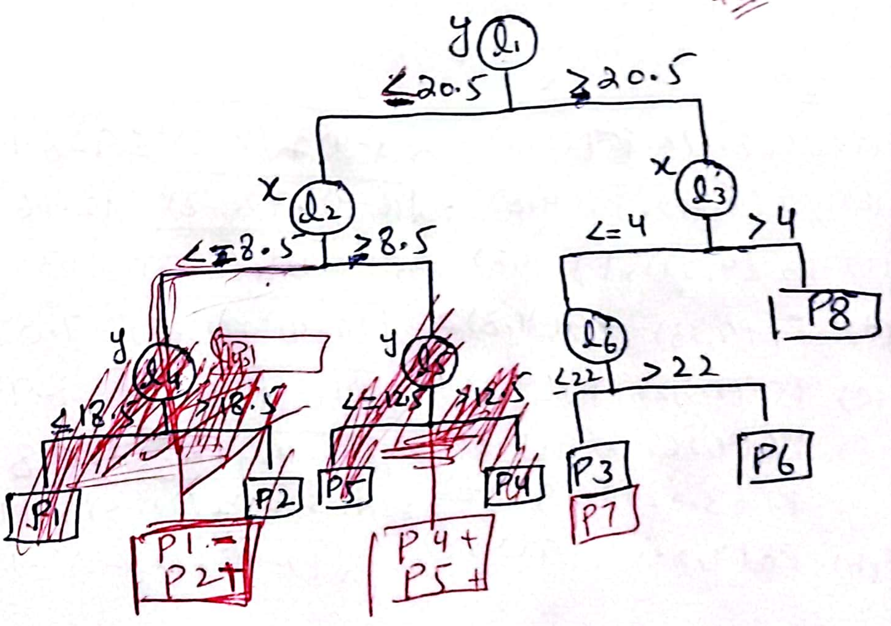
| | |
|---|---|
| 3 | 22 |
| 4 | 21 |
| 4 | 26 |
| 7 | 23 |

M=4

+5

KD Tree

+5



(iii)

~~Sorted on y~~

≤ >

+1

| | | | |
|---|---|---|---|
| 8 | 17 | 12 | 12 |
| 6 | 20 | 9 | 13 |

M 18.5

~~M=12.5~~

Sorted ≤ on y >

| | | | |
|---|---|---|---|
| 4 | 21 | 7 | 23 |
| 3 | 22 | | |
| 4 | 26 | | |

M = [illegible]

≤ >

| | | | |
|---|---|---|---|
| 4 | 21 | 4 | 26 |
| 3 | 22 | | |

b) Find the nearest neighbor for points P9(4,8) and P10(11,10) using the above constructed tree. Explicitly mention which branch you moved at each level.

① P9(4,8) ⇒ $l_1 - l_2 -$ P1 +2.5 so nearest neighbor is P1.
P2

② P10(11,10) ⇒ $l_1 - l_2$ – P5 +2.5 so nearest neighbor is P5.
P4

c) Classify the point P9(4,8) using 5 nearest neighbors using.

i) Majority voting approach

ii) Distance weighted majority weighting

(i)

Majority Voting

+5

(a) P1(8,17), P9(4,8) = $\sqrt{(8-4)^2+(17-8)^2} = 9.84$ - - .5

(b) P2(6,20), P9(4,8) = $\sqrt{(6-4)^2+(20-8)^2} = 12.16$ + - .5

(c) P3(4,21), P9(4,8) = $\sqrt{(4-4)^2+(21-8)^2} = 13$ - - .5

(d) P4(9,13), P9(4,8) = $\sqrt{(9-4)^2+(13-8)^2} = 7.07$ + - .5 +4

(e) P5(12,12), P9(4,8) = $\sqrt{(12-4)^2+(12-8)^2} = 8.94$ + - .5

(f) P6(4,26), P9(4,8) = $\sqrt{(4-4)^2+(26-8)^2} = 18$ .5

(g) P7(3,22), P9(4,8) = $\sqrt{(3-4)^2+(22-8)^2} = 14.03$ .5

(h) P8(7,23), P9(4,8) = $\sqrt{(7-4)^2+(23-8)^2} = 15.29$ .5

So 5 nearest neighbors are P1, P2, P3, P4, P5 +1 and according to majority Voting we can assign the class label to our test data P9(4,8) as " + ".

(ii) Distance weighted majority voting +5

distance of P9(4,8) with below datapoints and weighted distance

(a) P1 (8,17) = 9.84 = $\frac{1}{d^2} = \frac{1}{(9.84)^2} = 0.0103$ 0.25

(b) P2 (6,20) = 12.16 = $1/(12.16)^2 = 0.0067$ 0.25

(c) P3 (4,21) = 13 = $1/(13)^2 = 0.00591$ 0.25 +2

(d) P4 (9,13) = 7.07 = $1/(7.07)^2 = 0.02000$ 0.25

(e) P5 (12,12) = 8.94 = $1/(8.94)^2 = 0.01251$ 0.25

(f) P6 (4,26) = 18 = $1/(18)^2 = 0.00308$ 0.25

(g) P7 (3,22) = 14.03 = $1/(14.03)^2 = 0.00508$ 0.25

(h) P8 (7,23) = 15.29 = $1/(15.29)^2 = 0.00427$ 0.25

As 5 nearest neighbors are P1, P2, P3, P4, P5 and P1, P3 belong to class "-" and P2, P4, P5 to "+".

So,

(d)) for "-" +1

$[(9.84 * 0.0103) + (13 * 0.00591))] = 0.101352 + 0.07683$
$= \boxed{0.178182}$

For "+" +1

$[(12.16 * 0.0067) + (7.07 * 0.02000) + (8.94 * 0.01251))]$
$= 0.081472 + 0.1414 + 0.1118394.$
$= \boxed{0.3347114}$

+1

As the value for class label "+" is larger than "-" So we assign label "+" to our test datapoint P9(4,8).

***CLO # 2: Understand the nature of the data and apply data mining techniques to interpret the results.***

**Q3:** Consider the following ratings matrix with three users and six items. Ratings are on a 1-5 star scale. Compute the following from the data of this matrix. [5+10=15 marks]

| | | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |
|---|---|---|---|---|---|---|---|
| U1 | User 1 | 4 | 5 | 0 | 5 | 1 | 0 |
| U2 | User 2 | 0 | 3 | 4 | 3 | 1 | 0 |
| U3 | User 3 | 2 | 0 | 1 | 3 | 0 | 4 |

a) Treat missing values as 0. Compute the cosine similarity between each pair of users.

(i) user 1 and user 2

$\|U1\| = 16+25+0+25+1+0 = \sqrt{67} = 8.18$ $U1.U2 = 4*0+5*3+0*4+5*3+1*1+0*0$ +2.5

$\|U2\| = 0+9+16+9+1+0 = \sqrt{35} = 5.91$ $= 31$

$\cos(U1,U2) = \frac{U1.U2}{\|U1\|\|U2\|} = \frac{31}{8.18 * 5.91} = \boxed{0.6412}$

(ii) user2 and user 3 +1.5

$\|U2\| = \sqrt{35} = 5.91$

$\|U3\| = 4+0+1+9+0+16 = \sqrt{30} = 5.47$

$U2.U3 = 0*2+3*0+4*1+3*3+1*0+0*4 = 13$

$\cos(U2,U3) = \frac{U2.U3}{\|U2\|\|U3\|} = \frac{13}{5.91*5.47} = \boxed{0.4021}$

(iii) user 1 and user 3 +1

$U1.U3 = 4*2+5*0+0*1+5*3+1*0+0*4 = 23$ $\cos(U1,U3) = \frac{23}{(8.18)(5.47)} = \boxed{0.514}$

b) Treat the above given data for user1 and user2 as vector1 and vector2 with missing values treated as "0", apply a scaling factor of 2 and translation factor of 5 to vector2 and demonstrate and prove that cosine similarity is either invariant to scaling and translation or not.

vector1 = (4, 5, 0, 5, 1, 0)

vector2 = (0, 3, 4, 3, 1, 0)

→ scaling factor of 2 to vector 2

$vector2_s * 2 = (0, 6, 8, 6, 2, 0)$ +2

→ Translation factor of 5 to vector 2.

$vector2_t + 5 = (5, 8, 9, 8, 6, 5)$ +2

(i) $\text{Cosine}(V_1, V_2) \left(\frac{V1 \cdot V2}{\|V1\|\|V2\|}\right) = \boxed{0.6432}$ +1 (Already driven in part(a))

(ii) $(V_1, V_{s2})$

$\|V2\| = 0+36+64+36+4+0 = \sqrt{140} = 11.8321$ +2.5

$V_i \cdot V_{s2} = 4*0+5*6+0*8+5*6+1*2+0*0 = 62$

$$\cos\left(\frac{V1 \cdot V_{s2}}{\|V1\| \, \|V_{s2}\|}\right) = \frac{62}{(\cancel{6.7})(11.8321)} = \boxed{0.6405}$$

$(8.18)$

+2

(iii) $\|V_{t2}\| = 25 + 64 + 81 + 64 + 36 + 25 = \sqrt{295} = 17.1755$

$V_1 \cdot V_{t2} = 4*5 + 5*8 + 0*9 + 5*8 + 1*6 + 0*5 = 106$

$$\cos\left(\frac{V1 \cdot V_{t2}}{\|V1\| \, \|V_{t2}\|}\right) = \frac{106}{(\cancel{6.7})(\cancel{295})} = \cancel{0.00536} \; \boxed{0.7544}$$

$8.18 \times 17.1755$

+1

So, proved that ~~scaling factor~~ cosine similarity is invariant to scaling and not invariant to translation.

***CLO # 3: Evaluate the performance and effectiveness of different data mining models using appropriate matrices and validation techniques.***

**Q4:** Consider a binary classification problem where a model is trained to predict whether a patient has a certain medical condition (positive class) or not (negative class). The model is evaluated using a test dataset containing 100 patient records. The following confusion matrix is obtained from the model's predictions: [1*4 + 2+2+2 = 10 marks]

| | Predicted | |
|---|---|---|
| Actual | Condition | No condition |
| Condition | 30 TP | 10 FN |
| No condition | 5 FP | 55 TN |

Write appropriate formula and calculate the following:

a) No. of individuals classifier fails to identify who have the medical condition?

FN = 10 +1

b) No. of individuals classifier incorrectly identifies as having the medical condition?

FP = 5 +1

c) No. of individuals the classifier correctly identifies who do not have the medical condition?

TN = 55 +1

d) No. of individuals classifier correctly predicts that medical condition when they actually do have it.

TP = 30 +1

e) Accuracy

$$TP+TN \,/\, TP+FP+FN+TN \overset{+1}{=} \overset{+2}{\frac{30+55}{30+5+10+55}} \overset{+1}{=} 0.85$$

f) Specificity

$$TN \,/\, TN+FP = \overset{+2}{\frac{55}{55+5}} = 0.91$$

g) Precision

$$TP \,/\, TP+FP = \frac{30}{5+30} = 0.85^{+2}$$