עקרונות מדעי הנתונים

תרגיל מס' 2

תאריך ההגשה (דרך המודל): 09.12.2021.

שאלה 1

נתון המידע הבא:

x, z – משתנים מקריים בדידים. המשתנה x יכול לקבל ערכים a,b,c,d והמשתנה z יכול לקבל ערכים i, j, k.
בתוך הטבלה, הערך מציין את מספר המופעים בבסיס הנתונים של הערכים המתאימים.

data	z=i	z=j	z=k	Total
x=a	10	18	15	
x=b	2	10	20	
x=c	12	8	3	
x=d	12	17	9	
Total				

א. יש לחשב את האנטרופיה של המשתנים x ו-z.
ב. מהי האנטרופיה המקסימלית עבור כל אחד מהמשתנים x ו-z?
ג. יש לחשב אנטרופיה מותנית $H(z|x)$ וגם $H(x|z)$.
ד. יש לחשב Mutual Information (MI) עבור x ו-z. הוכיחו (ע"י חישוב) ש-MI סימטרי.

שאלה 2

מנהל חנות למוצרי בית החליט לנהל תוכנית מכירות חדשה ולשלוח פרסומות לגבי מספר מוצרים אך ורק ללקוחות עם פוטנציאל גבוה לרכוש את המוצרים. המנהל פנה לחברת מערכות מידע. בחברה החליטו לנתח את הנתונים של קניות העבר ולבנות עץ החלטה לכל סוג סוג של המוצר על מנת לסווג את הלקוחות לפי רמת פוטנציאל הקנייה (יקנה / לא יקנה).

הנתונים לגבי קנייה של מוצר ספציפי מכילים: רמת הכנסה של הלקוח, עונה שבה נשלחה פרסומת לגבי המוצר ללקוח, מחוז, סוג הבית, מס' קומות בבית והאם המוצר נקנה ע"י הלקוח.

להלן טבלה מרכזת בעניין:

ID	Income	Season	District	House Type	Num of floors	Bought?
1	High	Autumn	Suburban	Detached	1	No
2	High	Summer	Suburban	Detached	2	No
3	High	Spring	Rural	Detached	1	Yes
4	High	Winter	Urban	Semi-detached	1	Yes
5	Low	Spring	Urban	Semi-detached	1	Yes
6	Low	Autumn	Urban	Semi-detached	2	No
7	Low	Autumn	Rural	Semi-detached	2	Yes
8	High	Spring	Suburban	Terrace	1	No
9	High	Summer	Urban	Terrace	1	No
10	Low	Winter	Urban	Terrace	1	Yes
11	Low	Spring	Suburban	Semi-detached	1	Yes
12	Low	Autumn	Suburban	Terrace	2	Yes
13	High	Autumn	Rural	Terrace	2	Yes
14	Low	Spring	Rural	Detached	1	Yes
15	High	Spring	Urban	Terrace	2	No

השתמשו בנתונים אלו על מנת לבנות ולבדוק מודל ID3.

1. יש לחלק את הנתונים ל-training set (10 תצפיות הראשונות) ול-test set (5 תצפיות האחרונות).
2. בנו מודל ID3 על ה-training set (כאשר מספר מינימלי של תצפיות בעלה הוא 2).
3. חשבו את ה-Majority rule accuracy.
4. חשבו האם יש הבדלים מובהקים בין הדיוק של סט האימון והדיוק של סט המבחן (יש להשתמש ב-Normal Approximation to Binomial Distribution).
5. בדקו את המודל על ה-test set – האם קיבלתם overfitting?
6. לפי המודל שקיבלתם בסעיף 2, למי מהלקוחות החדשים הבאים כדאי לשלוח פרסומת?

ID	Income	Season	District	House Type	Num of floors	Bought?
16	High	Autumn	Suburban	Detached	2	?
17	High	Spring	Rural	Detached	2	?
18	Low	Spring	Rural	Semi-detached	1	?

7. ביצעו גיזום (post-pruning) לפי Pessimistic Error Pruning על המודל שקיבלתם בסעיף 2.

הוראות הגשה:

1. עליכם להגיש קובץ Word הכולל את מספרי תעודות הזהות של שני המגישים עם קו תחתון מפריד ביניהם בצורה הבאה: ID1_ID2.docx.
2. את הקובץ docx שיצרתם עליכם להעלות לתיבת ההגשה.
3. הגשה **בזוגות בלבד**.

ב ה צ ל ח ה ! !