

שאלות לדוגמה - מבוא למדעי הנתונים

אוסף שאלות לתרגול בלבד

הבהרה חשובה:

- אלה שאלות לדוגמה בלבד, ולא מבחן לדוגמה.
- מספר השאלות בבחינה יהיה שונה ממספר השאלות בקובץ זה.
- השאלות נועדו להמחיש את הנושאים ואת סוג החשיבה הנדרש, ולא את המבנה או את מספר השאלות בבחינה.
- משך הבחינה: שעתיים. אין חומר עזר.
- בכל שאלה אמריקאית יש תשובה נכונה אחת בלבד.

טופס תשובות

שאלה	תשובה	שאלה	תשובה
1		7	
2		8	
3		9	
4		10	
5		11	
6		12	

1. התבוננו בטבלה. איזה משפט מתאר בצורה הנכונה ביותר את איכות הנתונים?

num	age	gender	exercise_h	cats	dogs	gpa_grade	optimizm_lev	years_of_edu	weight	num_of_coffee
1	20	f	212	3	1	738	55	12	59.5	2
2	25	m	0	-2	0	624	83	2	190	2
3	22	m	6	1	3	603	50	11	78.1	5
4	23	m	2	0	2	707	62	15	69.1	2
5		f	1	0	1	735	78	12	70.6	2
6	21	f	8	0	2	636	42	11	71.8	0
7	18	m	8	2	0	645	51	15	59.5	5
8	24	f	4	2	3	571	74	15	67.8	2
9	21	m	0	1	2	632	64	15	74.2	5
10	23	m	13	0	1	613	75	14	70.6	2
11	24	f	5	2	3	584	56	12	71.4	2
12	28	m	3	2	1	532	100	12	69.8	0
13		f	8	3	2	663	31	12	63	1
14	22	m	7	3	2	653	76	15	63.8	1
15	21	m	10	2	2	593	52		64.3	3
16	20	m	n	3	4	680	73	15	66.1	4
17	21	m	15	2	2	439	61	10	68	5
18	21	m	0	2	2	605	69	12	66.4	2
19	23	m	4	2	2	687	65	10	71.4	1

טבלה מתוך נתוני סקר סטודנטים

א. הנתונים תקינים לחלוטין, משום שכל העמודות מופיעות בטבלה.

ב. קיימים רק ערכים חסרים, ואין בעיות נוספות.

ג. קיימים ערכים חסרים, ערך שלילי בלתי הגיוני, ערך שאינו מספר וערך חשוד מאוד.

ד. הבעיה היחידה היא שכל המשתנים בטבלה הם קטגוריאליים.

2. רוצים לבנות מודל לחיזוי SALARY מתוך המשתנים ID, NAME, DATE OF BIRTH, GENDER, CREDIT RATING ו-COUNTRY. איזו חלוקת תפקידים היא המתאימה ביותר?

א. SALARY הוא Target; ID ו-NAME הם Meta; המשתנים האחרים יכולים לשמש Features.

ב. ID הוא Target; SALARY ו-NAME הם Features.

ג. כל המשתנים, כולל ID ו-NAME, חייבים להיות Features.

ד. אין צורך להגדיר Target כאשר בונים מודל רגרסיה.

3. מהו הסדר הנכון של שלבי CRISP-DM?

א. הבנת העסק → הבנת הנתונים → הכנת הנתונים → מידול → הערכה → הטמעה.

ב. מידול → הטמעה → הבנת העסק → הכנת הנתונים → הערכה → הבנת הנתונים.

ג. הכנת הנתונים → הערכה → הבנת העסק → הטמעה → הבנת הנתונים → מידול.

ד. הבנת הנתונים → מידול → הבנת העסק → הערכה → הכנת הנתונים → הטמעה.

4. נתונה הסדרה: 2, 4, 7, 9, 10, 15. מהו החציון?

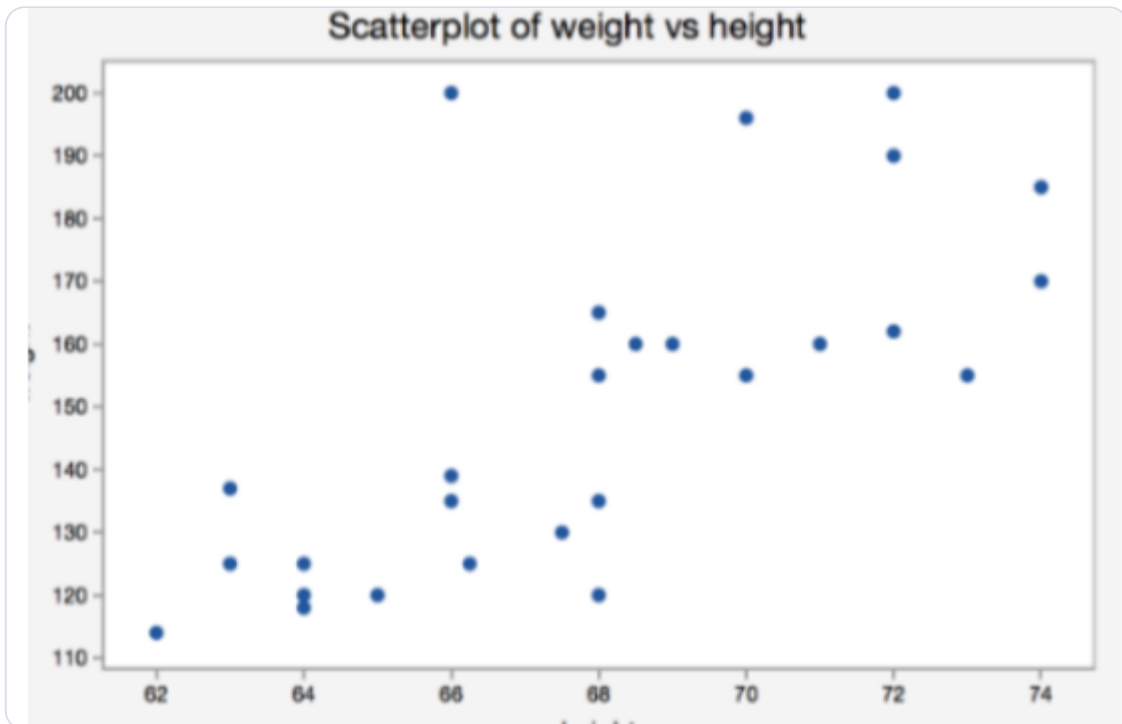
א. 7

ב. 8

ג. 9

ד. 47/6

5. התבוננו בתרשים הפיזור. איזה ערך של מקדם המתאם של פירסון הוא הסביר ביותר?



Scatterplot of weight vs height

א. $r \approx -0.85$

ב. $r \approx -0.10$

ג. $r \approx +0.60$

ד. $r \approx +1.00$

6. נתון מודל רגרסיה: $\text{Price} = 50 + 3 \cdot \text{Rooms} - 0.4 \cdot \text{Distance}$. כיצד מפרשים את המקדם 3 של Rooms?

א. הוספת חדר אחד מעלה את המחיר החזוי ב-3 יחידות, כאשר Distance נשאר קבוע.

ב. הוספת חדר אחד מורידה את המחיר החזוי ב-3 יחידות.

ג. כאשר מספר החדרים הוא 3, המחיר תמיד שווה ל-50.

ד. המספר 3 הוא מקדם המתאם בין Price לבין Distance.

7. בטבלה מוצגות תוצאות של שני מודלי רגרסיה. איזה משפט נכון?

מודל	R^2
מודל A	0.72
מודל B	0.41

א. מודל A עדיף לפי R^2 , והוא מסביר בקירוב 72% מהשונות של משתנה המטרה.

ב. מודל B עדיף, משום ש- R^2 קטן יותר תמיד טוב יותר.

ג. שני המודלים זהים, משום שבשניהם R^2 קטן מ-1.

ד. R^2 אינו מאפשר לומר דבר על איכות מודל רגרסיה.

8. שני Features במודל רגרסיה מתואמים זה עם זה ברמה של 0.98. מהי הבעיה האפשרית?

א. מולטיקולינאריות שעלולה להפוך את המקדמים ללא יציבים וקשים לפירוש.

ב. היעדר מוחלט של קשר בין המשתנים.

ג. המודל הופך אוטומטית למודל סיווג.

ד. כל הערכים החסרים מתמלאים באופן אוטומטי.

9. מהו השימוש העיקרי של רגרסיה לוגיסטית בקורס?

א. חיזוי ערך מספרי רציף בלבד.

ב. חיזוי ההסתברות להשתייכות למחלקה, למשל Yes או No.

ג. חישוב החציון של משתנה מספרי.

ד. חלוקת נתונים לקבוצות ללא משתנה מטרה.

10. בטבלה מוצגים ערכי AUC של שלושה מודלי סיווג. איזה משפט נכון?

AUC	מודל
0.84	מודל A
0.67	מודל B
0.50	מודל C

א. מודל A הוא הטוב ביותר מבין השלושה; מודל C דומה לניחוש אקראי.

ב. מודל C הוא הטוב ביותר, משום ש-AUC קטן יותר עדיף.

ג. כל המודלים שקולים, משום שאין AUC השווה ל-1.

ד. AUC מתאים רק למודלי רגרסיה עם Target מספרי.

11. מדוע לא נכון להעריך מודל רק על אותם נתונים ששימשו לאימון?

א. התוצאה על נתוני האימון עלולה להיות אופטימית מדי ואינה בודקת ביצועים על נתונים חדשים.

ב. משום שאסור להשתמש בנתונים מספריים בזמן אימון.

ג. משום שנתוני אימון משמשים רק לציור גרפים.

ד. אין כל הבדל בין הערכה על Train לבין הערכה על Test.

12. מה קורה ב-5-fold cross-validation?

א. מחלקים את הנתונים לחמישה חלקים; בכל סבב מאמנים על ארבעה חלקים ובודקים על החלק החמישי.

ב. מאמנים חמש פעמים על כל הנתונים ובודקים תמיד על אותם נתונים.

ג. משתמשים רק בחמישה משתנים מתוך כל הטבלה.

ד. מוחקים ארבע חמישיות מהנתונים ומשתמשים רק בחמישית אחת.

שאלות פתוחות קצרות

13. בניית וקטור Bag of Words לפי מילון נתון

אוצר המילים (המילון) מסודר כך: [data, science, orange, model, python].

עבור המסמך: "data science data model orange" — בנו את וקטור ה-Bag of Words לפי סדר המילים במילון.

שימו לב: הווקטור חייב לכלול רכיב לכל מילה במילון. אם מילה מהמילון אינה מופיעה במסמך, יש לרשום עבורה 0.

14. השוואה בין Train ל-Test

מודל רגרסיה קיבל $R^2_{\text{train}} = 0.94$ ו- $R^2_{\text{test}} = 0.46$.

מהי הבעיה האפשרית, ואיזו תוצאה משקפת טוב יותר את הביצועים הצפויים על נתונים חדשים? ענו עד שלוש שורות.

בהצלחה בתרגול!