

פתרונות - שאלות לדוגמה במדעי הנתונים

מפתח תשובות והסברים לשאלות התרגול

הבהרה חשובה: אלה פתרונות לשאלות לדוגמה בלבד, ולא למבחן לדוגמה. **מספר השאלות בבחינה יהיה שונה** ממספר השאלות בקובץ זה.

פתרונות לשאלות האמריקאיות

שאלה	תשובה	הסבר
1	ג	בטבלה מופיעים ערכים חסרים, $cats = -2$, האות n בעמודה מספרית, וכן $exercise_h = 212$ כערך חשוד מאוד.
2	א	SALARY הוא הערך שרוצים לחזות. מזהים כמו ID ושם אינם אמורים להשתתף כמאפייני חיזוי ולכן מתאימים לתפקיד Meta.
3	א	זהו הסדר הסטנדרטי של CRISP-DM. בפועל התהליך איטרטיבי וניתן לחזור לשלבים קודמים.
4	ב	יש שישה ערכים, ולכן החציון הוא ממוצע שני הערכים האמצעיים: $8 = 2 / (9 + 7)$.
5	ג	נראית מגמה חיובית ברורה אך לא מושלמת. לכן ערך חיובי בינוני-חזק, בערך 0.6, סביר יותר מ-1 או מערך שלילי.
6	א	ברגרסיה מרובת משתנים מפרשים כל מקדם כאשר שאר המשתנים נשארים קבועים.
7	א	R^2 גבוה יותר מצביע על שיעור גבוה יותר של שונות מוסברת. כאן מודל A מסביר כ-72% מהשונות.
8	א	מתאם גבוה מאוד בין מאפיינים עלול ליצור מולטיקולינאריות ולפגוע ביציבות ובפרשנות של המקדמים.
9	ב	רגרסיה לוגיסטית משמשת לסיווג ומחזירה הסתברות להשתייכות למחלקה.
10	א	AUC גבוה יותר מעיד על יכולת טובה יותר להבחין בין המחלקות. AUC של 0.5 הוא בקירוב רמת ניחוש אקראי.
11	א	

שאלה	תשובה	הסבר
		מודל עלול להתאים היטב לנתוני האימון אך להיכשל על נתונים חדשים. לכן משתמשים בקבוצת Test או ב-cross-validation.
12	א	בכל אחד מחמשת הסבבים חלק אחר משמש לבדיקה. כך כל תצפית משמשת פעם אחת לבדיקה ובשאר הסבבים לאימון.

פתרונות לשאלות הפתוחות

13. Bag of Words לפי מילון נתון

תשובה: [2, 1, 1, 1, 0].

כל רכיב בווקטור קשור למילה אחת במילון, לפי הסדר המדויק שבו המילים מופיעות במילון: data מופיעה פעמיים, science פעם אחת, orange פעם אחת, model פעם אחת, ו-python אינה מופיעה ולכן הרכיב שלה הוא 0.

גם מילה שאינה מופיעה במסמך נשארת חלק מהווקטור, משום שאורך הווקטור וסדר הרכיבים נקבעים לפי המילון ולא לפי המסמך הבודד.

14. Train לעומת Test

תשובה צפויה: הפער הגדול מצביע על התאמת-יתר אפשרית (overfitting). תוצאת ה-Test, $R^2 = 0.46$, משקפת טוב יותר את הביצועים הצפויים על נתונים חדשים.

בהצלחה בתרגול!