הטכניון
מכון טכנולוגי
לישראל

למידה עמוקה

הכנת הנתונים למידול

© kobymike@gmail.com

הכנת הנתונים למידול

- הנדסת מאפיינים - Feature engineering
- ניקוי והכנת הנתונים
- נרמול

הנדסת מאפיינים

© kobymike@gmail.com

הנדסת מאפיינים

- הוספת והורדת משתנים
  - על בסיס הבנת משמעות הנתונים
  - על בסיס קורלציה
- הפיכת משתנים קטגוריאליים למספריים
  - עם סדר
  - ללא סדר
- הפיכת משתנים מספריים לקטגוריאליים
- אינטראקציות ופולינומים

מושג	הסבר	דוגמה
אינטראקציות בין משתנים	יצירת תכונות חדשות על ידי הכפלה או שילוב של שני משתנים קיימים.	אם X_1 הוא כמות שעות למידה ו- X_2 הוא רמת קושי של מבחן, ניתן ליצור משתנה חדש: $X_{\text{new}} = X_1 \times X_2$.
פולינומים (חזקות של משתנים)	הוספת חזקות של משתנים קיימים כדי לאפשר למודל לזהות קשרים לא ליניאריים.	אם X הוא גיל התלמיד, ניתן להוסיף משתנים כמו X^2 (גיל בריבוע) ו- X^3 (גיל בחזקה שלישית).
מתי זה מועיל?	כאשר יש קשרים לא ליניאריים או אינטראקציות משמעותיות בין משתנים.	ברגרסיה ליניארית, הוספת מאפיינים כאלה יכולה לשפר את דיוק התחזיות. במודלים כמו Random Forest או XGBoost זה פחות נחוץ.

מחברת ללימוד:

https://colab.research.google.com/drive/1HQGJWUNgmb-McJaSuxIqH3pFK5KWq8yQ

תרגיל 1 – הנדסת מאפיינים

- טענו את בסיס נתונים titanic
  - (הערה: ניתן גם בסיסי נתונים אחרים לבחירתכם/ן)
- המטרה: חיזוי השרדות
- בנו מחברת וכתבו קוד בהתאם להנחיות הבאות:
  - אילו מאפיינים תכלילו בסיווג ואילו לא? מדוע?
  - האם אפשר לחשוב על בניית מאפיינים נוספים?
  - אילו מאפיינים קטגוריאליים נדרש להמיר למספריים? כיצד לעשות זאת?

ניקוי והכנת הנתונים

© kobymike@gmail.com

ניקוי והכנת הנתונים

- זיהוי ערכים חסרים
- זיהוי חריגים
- זיהוי כפילויות
- שיטות התמודדות

תרגיל 2 – ניקוי והכנת הנתונים

- המשיכו את פיתוח המחברת בהתאם לשאלות הבאות:
  - האם קיימים ערכים חסרים? אם כן טפלו בהם
  - האם קיימים ערכים חריגים? אם כן טפלו בהם
  - האם קיימות כפילויות? אם כן – טפלו בהם

נרמול

© kobymike@gmail.com

נרמול

/https://benalexkeen.com/feature-scaling-with-scikit-learn

נושא	הסבר
רקע	`Scikit-learn` היא ספרייה ללמידת מכונה ב-Python. אחת המשימות המרכזיות היא **סטנדרטיזציה** (התאמת הנתונים לממוצע 0 ושונות 1). בפעולה זו משתמשים ב-`()fit_transform` על נתוני האימון וב-`()transform` על נתוני הבדיקה.
מה עושה `()fit_transform`?	מחושב **הממוצע והשונות** של כל תכונה, והנתונים משתנים בהתאם. נועד **רק לקבוצת האימון**.
מה עושה `()transform`?	משתמש **בממוצע והשונות שנלמדו** מנתוני האימון וממיר בהתאם **את נתוני הבדיקה**.
למה זה חשוב?	אם נשתמש ב-`()fit_transform` גם על נתוני הבדיקה, המודל "יראה" אותם מראש ויהיה **מוטה** מה שיפגע בהערכתו על נתונים חדשים.
דוגמה בקוד	שימוש ב-`StandardScaler` כדי לאמן מודל תוך שמירה על עקרון ההפרדה בין נתוני האימון לבדיקה.
סיכום השיטות	`()fit_transform` – מחושב ממוצע ושונות ומבצע שינוי (**רק על נתוני האימון**). `()transform` – משתמש בממוצע והשונות שנלמדו ומבצע שינוי (**רק על נתוני הבדיקה**).
מסקנה	השימוש הנכון מבטיח שהמודל ילמד **רק מנתוני האימון** ויבדוק את עצמו **על נתוני הבדיקה כאילו הם חדשים לגמרי**.

תרגיל 3 – נרמול

- המשיכו את פיתוח המחברת בהתאם להנחיות הבאות:
  - בחנו את הסטטיסטיקה של הנתונים בעבודתכם.
  - נרמלו את הנתונים בשיטות שונות
  - איזו שיטה משיגה ביצועים טובים ביותר?
  - האם ניתן להסביר את שיפור הביצועים באמצעות שיטה זו?

שאלות?

© kobymike@gmail.com