למידה סטטיסטית וניתוח נתונים - תרגיל 1 (ניתוח בסיסי של מסד נתונים)
דדליין הגשה: יום רביעי, 22.3 ב-00:00

בתרגיל זה נתחיל לבצע בדיקות איכותיות בסיסיות של מסדי נתונים וקצת רגרסיה.

1. **בחירת מסד נתונים**: בחרו מסד נתונים שעשוי לעניין אתכם מהאתר Kaggle.

2. **תארו את מסד הנתונים שבחרתם**: מה המשתנים המסבירים בבעיה? הם סוגי המשתנים (למשל: קטגוריאלים, ממשיים וכו׳)? מהם המשתנים המוסברים? מה מנסים ללמוד?

3. האם אתם מצפים שתהיה תלות בין המשתנים המסבירים שלכם? נסו להשתכנע האם יש או אין תלות בין המשתנים והסבירו מדוע אתם חושבים כך. השתמשו בספריה של פייתון בשם corner באופן הבא: corner.corner(dataset), כאשר dataset זו מטריצת המשתנים המסבירים בלבד (אם יש יותר מדי, אפשר על מס׳ מוגבל שלהם – נניח, 10), והסבירו, לדעתכם מה מתארים הפאנלים הדו-מימדיים, ומה מייצגים הפאנלים החד-מימדיים. האם הם יכולים לעזור לנו להבין אם קיימת תלות בין משתנים מסבירים?

שלושת השאלות הבאות מתייחסות לקובץ הנתונים שניתן עם התרגיל - dataset.csv

4. השתמשו בספריה של פייתון בשם statsmodels, והפעילו רגרסיה לינארית על מסד הנתונים המצורף בקובץ התרגיל (dataset.csv) (העמודה הראשונה היא המסתנה המוסבר, והשאר - המשתנים המסבירים) - בעזרת השיטה של הספריה הנ״ל. הוציאו את וקטור הp-values עבור כל השיפועים (בטאות, זוכרים?). מה המשמעות של הp-values האלה?

5. כעת נרצה להשאיר רק את העמודות של המשתנים המסבירים הרלוונטיים. ממשו את שיטת הBackwards Elimination ע״פ מדד הR^2_adj, את שני אלו אתם בוודאי זוכרים מרגרסיה. אלו עמודות החליטה שיטה זו להשאיר? הסבירו. צרו plot (אחד של כולם, ולא מליון!) של כל אחת מהעמודות שנפסלה עם המשתנה המוסבר. מה רואים בו?

תזכורת: בשיטה זו נחשב תחילה את הR^2_adj עבור כל העמודות של X. לאחר מכן, נחסר עמודה עמודה, כאשר בכל פעם נפעיל מודל לינארי על העמודות הנותרות, נחשב את R^2_adj עבור מקרה זה ונבדוק האם הוא גבוה יותר מהערך הקודם. אם כן: עמודה זו ״מיותרת״ עבור המודל וניתן להתעלם ממנה.

הנחיות לכתיבת התרגיל: ניתן להגיש בזוגות (כאשר רק אחד מבני הזוג מגיש, ועל התרגיל מופיעים **שני מספרי הזהות ושמות המגישים**). ניתן להגיש בMarkdown וpdf (כולל קובץ הקוד במקרה השני).

שימו לב: אסתטיקה של ההגשה משנה ותעלה לכם בניקוד (ולא תהיה נתונה לשום ויכוח לאחר התרגיל הראשון). נא לכתוב בצורה מפורטת, להסביר ברור, לייצר את הגרפים באופן שניתן יהיה להבין לאן הם משויכים - רצוי עם כיתוב. להתאים גודל גופן של גרפים לטקסט (לא קטן מדי ולא גדול מדי), ולא להתפזר עם מס׳ הגרפים – אם אפשר, אחדו כמה רלוונטים לגרף בודד.

בהצלחה!