מעבדה בסטטיסטיקה 52568 - 2020-21, מטלה 10. להגשה והצגה ב-10.1

תיאור המשימה:

המעבדה עוסקת בחיזוי שינויים בדפוסי ההצבעה בשלוש מערכות הבחירות. יש להשתמש בקבצי תוצאות הבחירות על פי קלפיות בבחירות אפריל 2019, ספטמבר 2019, ומרץ 2020.

1. א. עבור כל זוג מבין שלושת הזוגות האפשריים של שלוש מערכות הבחירות מצאו את הקלפיות המשותפות לזוג והשתמשו בהן לאמידת מטריצת מעברי קולות. כאן ובכל השאלות השתמשו בשיטת nnls עם נרמול כל שורה לסכום 1 ואיפוס ערכים קטנים מ-0.5%, וכולל עמודה ושורה עבור "לא מצביעים". כללו בניתוחים את כל המפלגות שקיבלו לפחות רבע אחוז מהקולות בלפחות אחת מ-3 מערכות הבחירות (המפלגה הקטנה ביותר היא צומת). כלומר:
   - התאימו את מטריצת המעבר $M^{(ab)}$ עבור המעבר ממערכת הבחירות הראשונה למערכת הבחירות השניה.
   - התאימו את מטריצת המעבר $M^{(bc)}$ עבור המעבר ממערכת הבחירות השניה למערכת הבחירות השלישית.
   - התאימו את מטריצת המעבר $M^{(ac)}$ עבור המעבר ממערכת הבחירות הראשונה למערכת הבחירות השלישית.

   הציגו את שלושת המטריצות כ-heatmaps והסבירו את התוצאות

   ב. חשבו את השאריות הממוצעות עבור חיזוי מספר הקולות לכל מפלגה (כמו במעבדה 8 שאלה 4), והציגו ב-bar-plots. עבור אותן מפלגות בבחירות 2020, איפה מקבלים שאריות ממוצעות קטנות יותר, במטריצה $M^{(ac)}$ או במטריצה $M^{(bc)}$? הסבירו למה לדעתכם.
2. מצאו את קבוצת הקלפיות המשותפות ל-3 מערכות הבחירות. עבור כל קלפי ועבור כל אחד משני מודלי החיזוי על פי המטריצות $M^{(ab)}$ ו- $M^{(bc)}$ החוזים את ההצבעה בקלפי בבחירות מועד b ומעוד c בהתאמה, חשבו את השארית הריבועית הממוצעת של תחזיות המודל על פני המפלגות הגדולות (לפי שאלה 1) בקלפי זו ובמועד הבחירות המתאים.

   כעת הציגו scatter-plot של הנתונים בו כל נקודה מתאימה לקלפי ובכל ציר השגיאה הריבועית הממוצעת בקלפי במועד בחירות אחר. האם יש קשר בין טיב תחזיות המודל עבור הקלפי בין מערכות בחירות a,b לבין טיב התחזיות עבור מערכות בחירות b,c?
3. השתמשו בקלפיות המשותפות משאלה 2 והתאימו מודל לחיזוי מערכת הבחירות השלישית המשתמש גם במערכת הבחירות הראשונה וגם בשניה. כלומר בנו מודל רגרסיה בו הנתונים במטריצות $N^{(a)}, N^{(b)}$ מהווים את המשתנים המסבירים, והנתונים במטריצה $N^{(c)}$ את המשתנים המוסברים.

   א. חלקו את הקלפיות ל-train ו-test באופן אקראי כמו במעבדה 9. עבור מודל זה ועבור המודלים המשתמשים במטריצה $M^{(ac)}$ או במטריצה $M^{(bc)}$ חשבו את שגיאת ה-MSE הממוצעת על פני קלפיות ה-test. חזרו על התהליך (חלוקה ל-train-test באופן אקראי, אימון המודל וחיזוי) 10 פעמים ומצעו את שגיאות ה-test על פני 10 החזרות. איזה מודל משיג את שגיאת החיזוי הממוצעת הטובה ביותר?

   ב. אמדו כעת את המודלים על כל הנתונים (פעם אחת) וחשבו באמצעות שיטת ה-bootstrap את המקדמים הסיגניפיקנטיים השונים מאפס ברמת מובהקות 0.001 בכל אחד מ-3 המודלים. במודל המשתמש בשתי מערכות הבחירות, איפה יש יותר מקדמים סיגניפיקנטים שונים מ-0: במשתנים המסבירים של מועד a או של מועד b?

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- מותר להיות יצירתיים; נסו לחשוב על שיטות אחרות לחיזוי מועד c מתוך מועדי a,b.