מעבדה לסטטיסטיקה 52568 - 2019-20 ,מטלה 9. להגשה והצגה ב-5.1

תיאור המשימה:

המעבדה עוסקת בחיזוי שינויים בדפוסי ההצבעה בין שתי מערכות הבחירות וביצירת מדגם עבור מערכת בחירות. יש להשתמש בקבצי תוצאות הבחירות על פי קלפיות בבחירות מועד א ומועד ב ב-2019,

1. מיינו את 9782 הקלפיות המשותפות לשתי מערכות הבחירות על פי אחוז הקולות שהגיעו ל-6 המפלגות הגדולות הימניות הבאות: ליכוד, שס, יהדות התורה, ימין חדש, איחוד מפלגות הימין, זהות, מכלל הקולות הכשרים בכל קלפי בבחירות 2019א.
עבור פרמטר s שהוא מספר טבעי, נתאים מודל לחיזוי בחירות ב באמצעות בחירות א באופן הבא:
   - נחלק את רשימת הקלפיות הממוינת ל-s קבוצות שוות בגודלן ורצופות. כלומר למשל עבור s=2, קבוצה אחת תכיל את מחצית הקלפיות עם ההצבעה הגבוהה ביותר לימין, והקבוצה השניה את מחצית הקלפיות עם ההצבעה הנמוכה ביותר לימין. (אם מספר הקלפיות לא מתחלק ב-s נצרף את השארית לקבוצה האחרונה מבין s הקבוצות).
   - כעת נתאים לכל קבוצה בנפרד מטריצת מעבר M עבור המפלגות (ללא עמודת הלא מצביעים) בגודל 10*14 באמצעות שיטת nnls ללא נרמול השורות ואיפוס ערכים קטנים.

   עבור כל ערך s מ-1 עד 50, התאימו מודל זה בעזרת שיטת k-fold cross validation עם k=10 (עבור כל קלפי ב-test יש להשתמש במטריצה המתאימה מה-train לחיזוי על פי הקבוצה מבין s הקבוצות אליה שייכת קלפי זו). ציירו בגרף את שגיאת ה-MSE הממוצעת על פני k הקבוצות כפונקציה של s עבור ה-train וה-test (בשני צבעים שונים). דווחו על ערכי s הממזערים את שגיאת ה-train ואת שגיאת ה-test הממוצעות. במודל עבור איזה s הייתם משתמשים?
2. חזרו על שאלה 1 אבל הפעם חלקו את הקלפיות ל-s קבוצות בצורה אקראית . הסבירו כיצד השתנו תשובותיכם ומדוע
3. בחרו מדגמים בגדלים שונים 50 פעמים באופן הבא:
   - בכל פעם, עבור r=5,10,15,...,100 בחרו באקראי מדגם של r קלפיות בבחירות ספטמבר וחשבו את שכיחות ההצבעה ל-10 המפלגות הגדולות במדגם זה.
   - לאחר מכן חשבו עבור כל r את השגיאה הריבועית הממוצעת על פני כל המפלגות של תוצאות קלפיות המדגם לעומת שכיחות ההצבעה לכל מפלגה בתוצאות כלל הקלפיות.

   עבור כל r חשבו את הממוצע וסטיית התקן של השגיאה הריבועית הממוצעת של גודל מדגם זה בעזרת 50 המדגמים בגודל r שדגמתם. הראו בגרף error-bar את הממוצע וסטיית תקן של השגיאה הריבועית הממוצעת כפונקציה של גודל המדגם r.
4. (בונוס) הציעו שיטה אחרת לבחירת קלפיות המשיגה שגיאה ריבועית ממוצעת קטנה יותר משיטת הדגימה האקראית של שאלה 3, והשתמשו בה עבור אותם ערכי r בשאלה 3. הראו בגרף את ממוצע השגיאה הריבועית הממוצעת של שיטתם כפונקציה של r בהשוואה לשיטה האקראית משאלה 3 (בצבע אחר). מותר לכם להשתמש בתוצאות בחירות אפריל ובכל מידע אחר שאינו כולל את תוצאות בחירות ספטמבר בשיטת בחירת הקלפיות. שיטתכם יכולה להיות דטרמיניסטית או מערבת אקראיות

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- מותר לכם להיות יצירתיים; נסו לחשוב על שיטות אחרות לאמידת מודל ולשפר את ה-MSE על ה-test לעומת השיטה שבשאלה 1.