**מבוא למערכות לומדות – תרגיל בית 2**

מגישים:

**שם:** אסף חאייק ברוך **ת.ז:** 206783441

**שם:** בן בנוז **ת.ז:** 207570573

**הסבר על פתרון הבעיות**

כדי לפתור את הבעיות שהוצגו בתרגיל השתמשנו במסווגים שפותרים את בעיית קביעת ההצבעה של בוחר לפי הנתונים עליו (הבעיה הבסיסית), לכן לכל בעיה הגדרנו מעטפת שתהיה אחראית על מסווג סטנדרטי, הגדרנו לכל בעיה גם פונקציה לשערוך טיב הפלט של מודל שפותר בעיה.

1.המנצח בבחירות

החלטנו שפתרון לבעיה זו יכול להיגזר מפתרון לבעיה השנייה של התפלגות הקולות בין המפלגות, ולכן החלטנו להשתמש פחות או יותר באותו הפתרון רק עם פונקציית הערכה שונה, פונקציית ההערכה במקרה הזה מתייחסת להיסטוגרמת הקולות של המפלגות כ-logits של softmax כך שטעויות בחיזוי אצל המפלגות המובילות הרבה יותר משפיעות על הערכת טיב החיזוי, אנחנו מחשבים את ה"מרחק" בין ההיסטוגרמה האמיתית של התפלגות הקולות לבין זאת שחזינו (ה-softamx של ההיסטוגרמות) ומעדיפים ערכים יותר קרובים ל-0.

2.תוצאות הבחירות

החלטנו לנסות לחזות היסטוגרמה של מספר הקולות לכל מפלגה באופן הסתברותי.

כלומר לכל מצביע חישבנו את ההסתברות שהוא יצביע למפלגה כלשהי והערך בהיסטוגרמה לכל מפלגה היה סכום ההסתברויות האלו אצל כל המצביעים, בדומה לבעיה הקודמת אנחנו מעריכים את טיב החיזוי כמרחק בין ההיסטוגרמה האמיתית לזאת שחזינו רק הפעם בלי softmax.

3.מצביעים סבירים לכל מפלגה

בדומה לתהליך שביצענו בפתרון הבעיות הקודמות חישבנו את ההסתברות של כל מצביע להצביע לכל מפלגה והחזרנו לכל מפלגה רשימה של מצביעים שיש להם הסתברות גבוהה מערך שהמשתמש הגדיר (threshold) להצביע לה, אנחנו בחרנו להשתמש בערך דיפולטי של 60% .

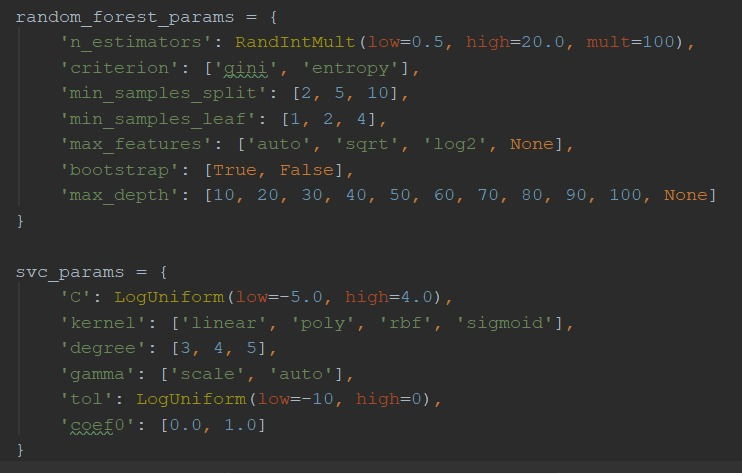
פונקציית ההערכה במקרה זה הינה ממוצע של ניקודי הf1 בין ההמצביעים שקבענו שסביר שיצביעו למפלגה מסויומת לבין האנשים שבאמת הצביעו לאותה מפלגה (ממוצע על כל המפלגות).

**המודלים שבחרנו לעבוד איתם**

החלטנו לבחור מודל לינארי ומודל של עץ החלטה ולפי צרכי הבעיות היינו חייבים שלאותם מסווגים תהיה פונקציה שמחזירה הסתברות לכל סיווג. בגלל מגבלות אלה ואיכות המסווגים החלטנו להשתמש ב-svc וב Random Forest .

כיוונון היפר פרמטרים

בחרנו לכוונן היפר פרמטרים לכל בעיה לפי חיפוש אקראי על מרחב פרמטרים שהגדרנו.

לכל פרמטר שבמרחב הגדרנו התפלגות ערכים אקראית שתעבוד עם RandomizedSearchCV שעושה חיפוש אקראי על היפר פרמטרים עם  
k-Fold cross validation לפי פונקציות ההערכה שהגדרנו. גם ווידאנו שהחיפוש נעשה עם אותו מצב אקראי לכל מסווג. ההתפלגויות שבחרנו היו בטווחי ה-log והמכפלות מכיוון שהיה לנו חשוב יותר למצוא את סדר הגודל של ההיפר פרמטרים מאשר הגודל המדויק.

בחרנו להשתמש בחלוקה לשלושה חלקים ו-10 ניחושים אקראיים.

בחירת המודל האופטימלי לבעיה

לאחר שכיווננו את הפרמטרים של המסווגים שלנו אימנו אותם על כל סט האימון והשווינו את החיזוי שלהם על סט הוולידציה והערכנו את התוצאות לפי פונקציות ההערכה שלנו.

לאחר מכן פשוט בחרנו את המסווג עם הניקוד הכי גבוה.