מעבדה 5 – מסווג לינארי מסוג רגרסיה לוגיסטית

כפתרון למעבדה זאת יש להגיש מחברת Jupyter עם פתרונכם. לאחר הרצתה במלואה הגישו אותה בשתי גרסאות:

א. כקובץ ipynb כשהוא מוכן להרצה מחדש
ב. כקובץ pdf (על קובץ זה להיות זהה בכל פרטיו לקובץ בסעיף א, פרט להיותו בפורמט שונה)

בהצלחה!

במעבדה זאת תקבלו את הדאטה ואת ערכי הפרמטרים של מסווג לינארי, ותתבקשו לחשב את דיוק המסווג ולהציג את התוצאות (שימו לב כי תרגיל זה **אינו** כולל את אלגוריתם הלמידה עצמו).

חלק א – מסווג לינארי

1. צרו מחברת בשם LinearClassifiers.ipynb וטענו אליה את המידע הנתון בארבעת הקבצים הבאים:
המאפיינים של סדרת האימון מהקובץ Xtrain.txt
הסיווגים של סדרת האימון מהקובץ Ytrain.txt
המאפיינים של סדרת המבחן מהקובץ Xtest.txt
הסיווגים של סדרת המבחן מהקובץ Ytest.txt
(מרחב המאפיינים הוא דו-ממדי, והתיוגים בינאריים)
2. הציגו את הדוגמאות בגרפים דו-ממדיים מתאימים (ראו דוגמא בנספח).
3. טענו את וקטור הפרמטרים של המסווג מהקובץ Coefficients.txt והדפיסו את ערכם.
4. ממשו פונקציה בשם Classify אשר
   - מקבלת כקלט את וקטור הפרמטרים של המסווג הלינארי, ואוסף של דוגמאות לא מסווגות
   - מחזירה כפלט את Ypredicted – וקטור עם סיווגן של הדוגמאות הנ"ל, כפי שחושבו ע"י המסווג.
5. ממשו פונקציה בשם accuracy אשר
   - מקבלת כקלט את וקטור הסיווגים האמיתיים Y, ואת וקטור הסיווגים שחושבו ע"י המסווג Ypredicted
   - מחזירה כפלט את דיוק המסווג, כלומר את אחוז הדוגמאות המסווגות נכון, עבורן Y=Ypredicted
6. תוך בחינת הדאטה ווקטור הפרמטרים של המסווג הנתון, נסו למצוא וקטור פרמטרים אחר שישיג ביצועים טובים יותר בסיווג המידע (של סדרת האימון, סדרת המבחן, או שתיהן). תארו במספר משפטים את הכיוונים שבחנתם בניסיונותיכם ואת התוצאות שקיבלתם. בהמשך השאלה השתמשו בוקטור החדש.
7. חשבו והציגו את מטריצות הערבול (confusion matrix) של המסווג הנ"ל על סדרת האימון ועל סדרת המבחן (שימו לב שיש לחשב מטריצה נפרדת לכל סדרה)

חלק ב – רגרסיה לוגיסטית

8. ממשו פונקציה בשם ProbabilisticLogRegClassifier אשר
   - מקבלת כקלט את וקטור הפרמטרים של המסווג הלינארי, ואוסף של דוגמאות לא מסווגות
   - מחזירה כפלט את P(Ypredicted=1) – וקטור הכולל את ההסתברויות שסיווגי כל אחת מהדוגמאות הנ"ל הינו 1, כפי שחושבו ע"י מסווג מסוג Logistic Regression ע"פ הנוסחה

$$P(y=1|\mathrm{x},\mathrm{w},\mathrm{w}_0)=\frac{1}{1+e^{-(w^Tx+w_0)}}$$

9. ממשו פונקציה בשם FinalClassification אשר
   - מקבלת כקלט את הוקטור P(Ypredicted=1), וערך סף $0 \le th \le 1$
   - מחזירה כפלט את הסיווג הסופי של כל אחת מהדוגמאות

$$y=\begin{cases}1 & \text{if } P(y=1|\mathrm{x},\mathrm{w},\mathrm{w}_0)>th\\ 0 & \text{otherwise}\end{cases}$$

10. חשבו והציגו את מטריצות הערבול (confusion matrix) של המסווג הנ"ל על סדרת האימון ועל סדרת המבחן (שימו לב שיש לחשב מטריצה נפרדת לכל סדרה). השוו אותן לאלו שקיבלתם בשאלה 7, ודונו בקצרה בתוצאות.

נספח

להלן רעיון אפשרי לויזואליזציה של חלק מהמידע והתוצרים בחלק א, אך מומלץ מאוד להביא לידי ביטוי את היצירתיות שלכם במציאת דרכים נוספות להציג את המידע והממצאים שהתבקשתם לחשב ולהציג.