

מעבדה 7 – מעקב אחרי אימון מסוג logistic regression

שימו לב:

- ניתן ואף מומלץ להשתמש בקוד שמימשתם במעבדות קודמות במידה והוא יכול לסייע לכם בפתרון המעבדה.
- בתרגיל זה תקבלו את הדאטה ואת ערכי פרמטרי המסווג עבור מספר עדכונים שלהם ע"פ אלגוריתם gradient descent, ותידרשו לחקור באמצעותם מספר היבטים של תהליך האימון. תרגיל זה אינו כולל את אלגוריתם עדכון הפרמטרים עצמו.

במעבדה זאת תמדדו את השתנות ביצועיו של מסווג רגרסיה לוגיסטית במהלך אימונו עם אלגוריתם gradient descent.

1. צרו מחברת jupyter חדשה בשם LogisticRegressionTrainingMonitoring.ipynb וטענו אליה את המידע בקבצים 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
- וקטור המשקלים של המסווג באתחול ולאחר כל אחת מ-300 האיטרציות של תהליך האימון מהקובץ Coefficients.txt

2. לכל m (אינדקס איטרציית עדכון הפרמטרים, המצוין בעמודה הראשונה בקובץ Coefficients.txt), חשבו עבור סדרת האימון את הגדלים הבאים והציגו אותם בגרפים מתאימים (הניחו $\text{threshold}=0.5$ בכל מקום בו נדרש ערך הסף)

- דיוק המסווג $P_c(f, \mathcal{D}_{\text{train}})$ (הציגו המידע בגרף עם ערכי m בציר האופקי)
 - ערך פונקציית ה cross-entropy $\mathcal{L}_{\text{CE}}(\mathcal{D}_{\text{train}}, w^{(m)})$ (הציגו המידע בגרף עם ערכי m בציר האופקי)
 - עבור ערך ה- m הגדול ביותר בקובץ הפרמטרים, חשבו והציגו את דיאגרמת ROC ומטריצת הערבול confusion matrix של המסווג שהתקבל
- דונו בתשובתכם בקצרה בתוצאות שהצגתם.

3. חזרו על שאלה 2 עבור סדרת המבחן (כלומר חשבו את $P_c(f, \mathcal{D}_{\text{test}})$, ערך פונקציית ה cross-entropy $\mathcal{L}_{\text{CE}}(\mathcal{D}_{\text{test}}, w^{(m)})$, ועבור ערך ה- m הגדול ביותר גם את דיאגרמת ROC וה- confusion matrix). דונו בתשובתכם בקצרה בתוצאות שהצגתם תוך השוואתם לתוצאות בשאלה 2.

לשימושכם, להלן מספר רעיונות אפשריים כיצד ניתן להציג גרפית חלק מהמידע הנ"ל (עם דאטה שונה מזה שקיבלתם בתרגיל). ניתן להשתמש בהם, אך בנוסף מומלץ מאוד להביא לידי ביטוי את היצירתיות שלכם במציאת דרכים נוספות להציג את הממצאים שהגעתם אליהם, ואת ההבנה שלכם אותם.



