מבוא ללמידת מכונה

תרגיל 3

תאריך הגשה: באתר הקורס

<u>שאלה 1</u>

העזרו במצגת שהועברה בקורס וממשו מחלקה הנקראת LogisticRegression שתכיל בנוסף לשאר גם את המתודות fit, predict, score:
fit(self, X, y) – מקבלת את X מטריצה המכילה את ה features, y מכילה את הlables, ומאמנת מודל בעזרת אלגוריתם רגרסיה לוגיסטית עם גראדינט דיסנט. את וקטור המשקולות w שמרו בשדה _weights של המחלקה כך שתהיה אפשרות לגשת אליו.

predict(self, X) - מחשבת את הפלט של המודל על הדוגמאות ב-X.

predict_proba(self, X) - מחשבת את ההסתברויות שהמודל מתאים לדוגמאות ב-X.

score(self, X, y) - מחשבת את אחוז הדוגמאות ב-X שהרשת המאומנת מסווגת בצורה נכונה, מחזיר תוצאה בין 0 ל-1.

<u>שאלה 2</u>

ממשו פונקציית main שמשתמשת במחלקה LogisticRegression שכתבתם על ה- ham/spam dataset שמופיע בmama. המטרה היא להעריך את ההסתברות שהודעת מייל הינה spam.

הדפיסו את התוצאות ואת וקטור המשקולות בסוף התהליך.

שימו לב: יש לעשות עיבוד מקדים על הdataset כך שתוכלו להשתמש בו.

<u>שאלה 3</u>

כתבו פונקציה לשרטוט ה ROC curve עבור המחלקה LogisticRegression על הham/spam dataset.

בשימוש במחלקה שלכם לצורך סיווג דואר זבל, איזה threshold הייתם בוחרים לצורך תרגום ההסתברויות לסיווג?

<u>שאלה 4</u>

א. הוסיפו למחלקה LogisticRegression שלכם אפשרות ל-multiclass classification. השתמשו ב- one vs rest לצורך כך.
ב. ממשו פונקציית main שמשתמשת במחלקה LogisticRegression שכתבתם על ה- iris dataset. שימו לב שבדאטהסט הזה יש יותר משתי מחלקות.

אפשר להוריד את הדאטהסט מ-sklearn, בעזרת הקוד הבא:

```
from sklearn import datasets

iris = datasets.load_iris()
```

תזכרו: בעבודה על הדאטהסטים השונים (ham/spam, iris) בחירת הייצוג תמיד בשליטתכם. בין הכלים שניתן להשתמש לצורך הייצוג: כלים שאת מכירים (PCA, scaling, polynomial features, ועוד), וכל רעיון יצירתי אחר שיש לכם.