

דו"ח תרגיל בית 1

- קיבלנו דאטה בקובץ טסקסט המכיל שורות של 9 פיצ'רים, בערכים שלמים בין 1 ל 3, וקבוצה אליה שייכים אותם ערכים של הפיצ'רים "g1" ו "g2".
חסר מידע על הדאטה, מהי כל קבוצת השתיכות, מה המשעות של הסיווג בין הקבוצות ומה כל נתון של כל פיצ'ר מציין.
- עשיתי מספר התאמות בקוד על מנת לענות על הגדרות התרגיל וע"מ להתאים את המודל לדאטה החדש שנכנס:
 - פונקציה הקוראת את הנתונים מהקובץ (ויכולה גם לערבבם), ממירה את שתי הקבוצות שבקובץ ממחרות למספר ע"מ שנוכל לעבוד עם הפרספטרון, קבוצה "g1" הומרה ל 0 וקבוצה "g2" ל 1.
 - בנוסף הפונקציה מחלקת את רשימת הנתונים לנתונים לאימון ונתונים לבדיקה לפי היחס הנבחר ומוסיפה לכל רשימת מידע גם את אינפוט הביאס.
 - התאמה של גודל האינפוט לנורון, כך שיתאים ל 9 פיצ'רים בנוסף לביאס.
 - הוספתי פונקציה שמחשבת את הloss, בשיטת ה MSE, לאחר כל איטרציית אימון ומציגה את קצב התקדמות של ה loss אחרי כל epoch.
 - חישוב של ה recall של המודל והדיוק (precision) שלו.
- שיטת העבודה:
 - טעינת הדאטה והכנסתו למודל לקראת העבודה עם הפרספטרון, אני בחרתי גם לערבב אותו כדי להגדיל את הסיכויים להגיע למציאות יותר "אמיתית".
 - חלוקת הדאטה לאימון ובדיקה
 - יצירת הפרספטרון עם גודל אינפוט מותאם, learning rate ו שלו.
 - הרצת המודל לאימון והדפסת נתוני האימון המשקולות וגרף של התקדמות ה loss.
 - בדיקת המודל מול דאטה חדש שלא היה באימון והדפסת התוצאות.
 - בדיקת המודל מול הדאטה שהוא אומן איתו והדפסת התוצאות
 - בחרתי לעשות זאת מאחר ולדעתי אם התוצאה כאן לא תהיה טובה זה מעיד שהמודל לא יציב (ללא קשר לאיכות החוזי), מאחר שאפילו על מה שהוא כבר מכיר הוא אינו משיג תוצאה טובה – אז זה מעיד על איכות המשקולות.
- המסקנות שלי לגבי המודל:
 - שיחקתי יותר עם כמות המודל לאימון, כמות ה epochs שהמודל מתאמן וה learning rate, אחרי כמה משחקים ראיתי שבסוף אני מגיע לתוצאה מסויימת שלא מתשתפרת כ"כ שאני מנסה להקטין את ה learning rate להגדיל את כמות המידע לאימון ואת כמות ה epochs. לכן בחרתי את המספרים שהניבו לי את התוצאה האופטימלית בכמות זמן והחישוב הכי פחותה.
 learning rate: 0.0000000001
 Epochs: 10,000
 ו 80% מכל הדאטה שומש לאימון.
 - נראה שהמודל לא מצליח כ"כ להפריד את הדאטה, למעשה לא הגעתי למודל שהוא יציב דיו בשביל שאוכל לסמוך על החוזי שלו, אך גם מצד שני לא מצאתי עם מה איזה מספרים אני יכול לשחק בשביל להניב תוצאות טובות יותר.
 התוצאות דיוק אליהם הגעתי היו לכל היותר באזור ה 70% ולרוב נעו לכיוון ה 65%.

- גם לגבי התוצאות של ה recall לא זהיתי כיוון מובהק של תוצאות, התוצאות היו מאוד מפוזרות ונעו בין אחוזי הצלחה גוברים ולאחוזים אפסיים.
- ראיתי ש loss שלי לא מתכנס לכיוון ה 0 אלא נשאר באזור ה 0.35, גם כאשר ניסיתי לשחק עם כמות ה epochs, כשהגלתי כמות הדאטה לאימון וגם כשהקטנתי והגדלתי את ה learning rate.
- ניתן לראות שאחוזים של ה recall וה precision של המודל על הדאטה שהוא אומן עליו וגם על הדאטה החדש, מה שמעיד המודל אכן יציב אך לא מנתפק תוצאות חוזי מספיק טובות (כמובן ביחס לכמות הדאטה הקיים)
- בשביל להגיע לתוצאות טובות יותר לדעתי צריך קודם כל לקבל יותר דאטה ויותר מידע על הדאטה. ככה שנוכל אולי לשחק איתו (לנרמל אותו) בצורה שתוכל להטיב אם המודל הנוכחי והדרך שבה הוא ממש את ה activation function וה loss function.
- אני חושב שגם צריך לשנות את ה activation function, שתביא לידי ביטוי טוב יותר את תוצאת הפרספטרון.

למטה ניתן לראות תוצאה אחת מבין הריצות, כאן כפי שציינתי העדפתי לתת 80% מהדאטה לאימון, ו 20% לבדיקה. המודל ביצע epochs 10,000 בקיצב לימוד של 0.0000000001

```
=====
Loading Data Process:
=====
Total Training Data: 766 (79.96% of dataset)
Total Testing Data: 192 (20.04% of dataset)
=====
Training Process:
=====
Total Epochs: 10000
Model Loss: 0.34856396866840733
Model Weights: [0.000267 0.000492 0.000472 0.000487 0.00046 0.000502 0.000474 0.000482
0.000483 0.000486]
=====
Checking the model on the new data
Checking Process:
=====
Total Checking Inputs: 192
Total True Positive: 0
Total False Negative: 65
Total Correct Predictions: 127
Model Recall: 0.0%
Model Precision: 66.15%
=====
Checking the model on the trained data
Checking Process:
=====
Total Checking Inputs: 766
Total True Positive: 0
Total False Negative: 267
Total Correct Predictions: 499
Model Recall: 0.0%
Model Precision: 65.14%
```

