מבוא לבינה מלאכותית, 236501

מטלה 3 – עצי החלטה ולמידת מכונה

מגיש: שהם אלחנתי, 208501684

שאלה 1

1. -
2. 

שאלה 2

הטענה נכונה.

נרמול של ערכים לטווח יביא לנרמול של ערך הסף, ה-Threshold, לבחירת כל תכונה. חישוב האנטרופיה ותוספת המידע בכל שלב מתבצעים על פי **הסיווג** – **שאינו עובר נרמול**. אי לכך, נרמול אינו משפיע על הדיוק ונקבל את אותה התוצאה.

שאלה 3 – גיזום מוקדם

1. התאמת יתר (overfitting) היא מצב בו הקטנת שגיאת האימון – מגדילה את שגיאת המבחן, ונובעת בדרך כלל מדוגמאות רועשות.

גיזום עצי החלטה נעשה כדי להקטין את העץ, ולהחליש את אפקט התאמת היתר. בגיזום מוקדם, עוצרים את גידול העץ על אף שהעלה אינו אחיד.

לפיכך גיזום מגדיל את שגיאת האימון בתקווה להקטין את שגיאת המבחן.

1. –
2. בחרתי את הערכים: לפרמטר . הדיוק הממוצע עבור כל ערך הוא כדלהלן:

**להוסיף תוצאות!!!!**

מצורף גרף המציג את השפעת הפרמטר על הדיוק:

ניתן לראות כי עבור , ערכי הדיוק הם זהים והגבוהים ביותר, ואילו עבור ערכים הירידה בדיוק חדה מאוד.

*לפיכך עבור מתקבל הגיזום הטוב ביותר, עם ערך .*

1. ***לימוד הכל***

*שאלה 4*

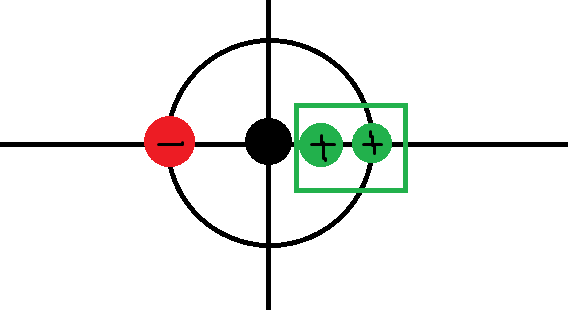
1. *על מנת לשפר את ערך ה- נפעל בצורה הבאה: כאשר מגיעים לכמות דוגמאות נמוכה (למשל 10), לא נבחר בתיוג בעל הכמות הגבוהה ביותר, אלא ניקח מרווח. כלומר, אם הכמות היא כחצי-חצי לטובת ,* ***נבחר ב-*** *מתוך הנחה שייתכן והאבחנה שגויה והובילה לטעויות, ועדיף להתחשב ב-.*

*בפועל, הדבר משפר גם את הדיוק.*

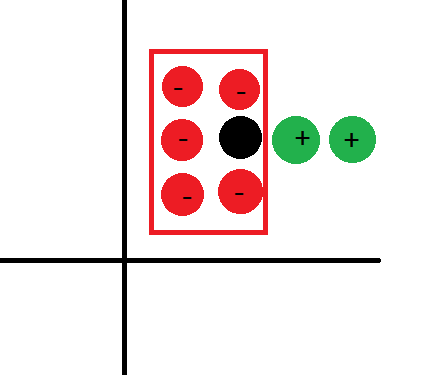
*שאלה 5*

*בכל הדוגמאות הבאות, ירוק ואדום הן דוגמאות האימון, כאשר ירוק זה חיובי ואדום שלילי, ושחור היא נקודת המבחן.* ***כמו כן, כל הגרפים מתארים סיווג שהוא מסווג המטרה.***

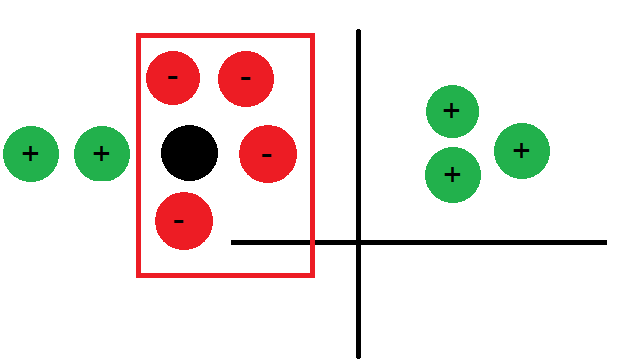
1. *( צודק, טועה) עבור ו- השחור יסווג כירוק, ועבור בגלל שיש שתי נקודות במרחקים זהים, השחור יסווג גם כירוק כי ערך שלו גדול יותר, לכן יטעה עבור כל .*

**

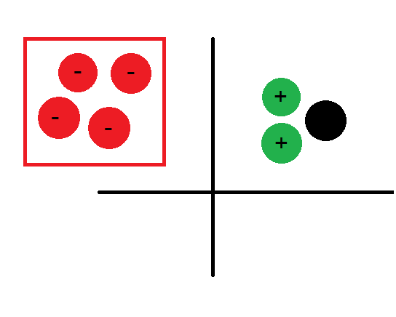
1. *( צודק, טועה) הסיווג של הנקודה השחורה הוא ירוק, אך יסווג אותה כאדומה. עבור ו-, יסווג אותה כירוקה.*

**

1. *(שניהם טועים) הסיווג הנכון של הנקודה השחורה הוא ירוק, אך מסווג אותה כאדומה. עבור לדוגמא, גם יסווג אותה כאדומה.*

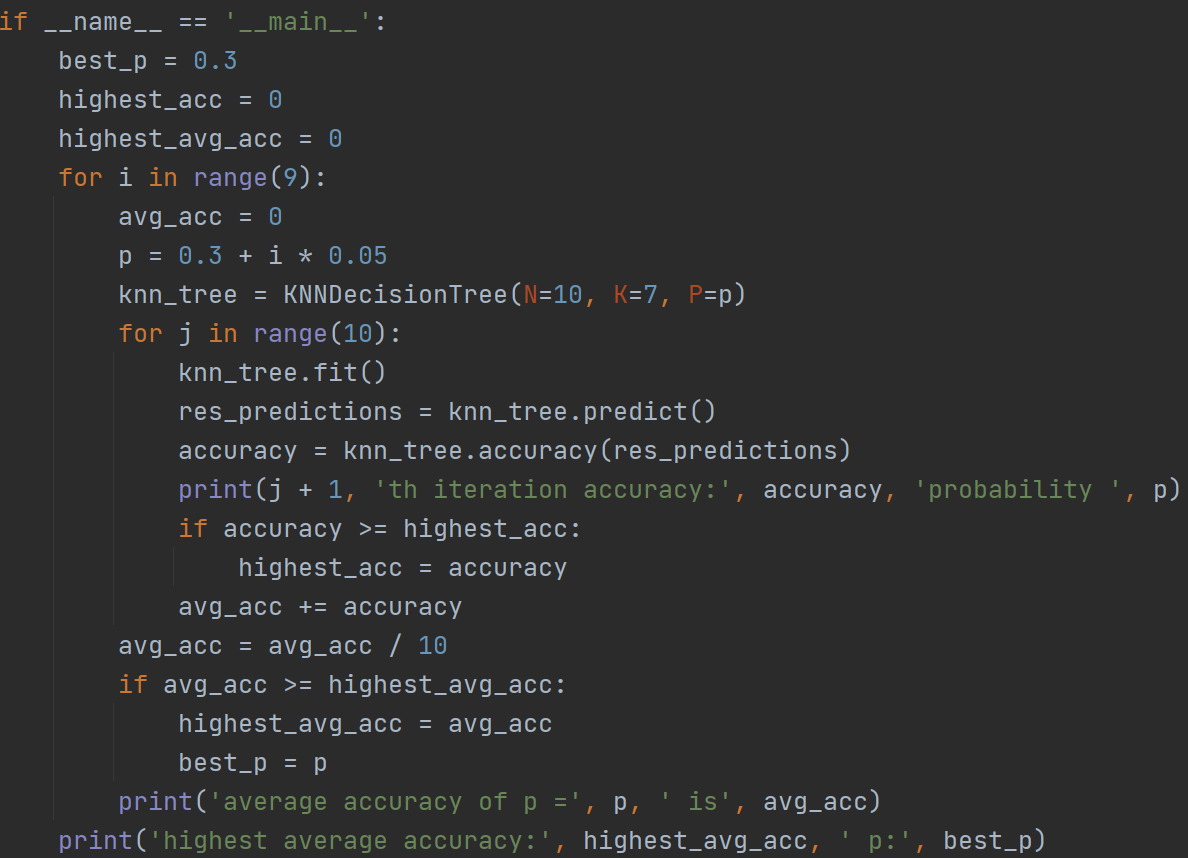
**

1. *(שניהם צודקים) עבור , יסווג את הנקודה השחורה כירוקה, ולכן יתקבל מסווג המטרה כמו , ושני האלגוריתמים צודקים עבור הנקודה השחורה.*

**

שאלה 6

בחרתי באופן שרירותי , ובדקתי לכל בקפיצות של , מתי אקבל את הדיוק הממוצע הגבוהה ביותר. לשם הבדיקה הרצתי את הקוד הבא:

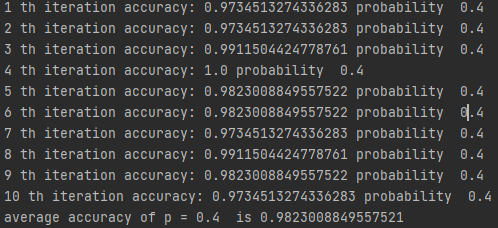


להלן טבלה מסכמת לערכי והדיוק הממוצע שהתקבל:

|  |  |
| --- | --- |
|  |  |
| 0.9805309734513272 | 0.3 |
| **0.9849557522123893** | **0.35** |
| 0.9823008849557521 | 0.4 |
| 0.9796460176991149 | 0.45 |
| 0.9805309734513272 | 0.5 |
| 0.9796460176991151 | 0.55 |
| 0.9752212389380531 | 0.6 |
| 0.9787610619469026 | 0.65 |
| 0.9752212389380531 | 0.7 |



הדיוק הממוצע הגבוה ביותר התקבל עבור . אמנם היו ערכי אחרים עבורם התקבל דיוק נקודתי גבוה יותר (לדוגמא 1 עבור כפי שרואים בצילום הבא:

),

אך הממוצע עבור היה גבוה יותר, ולכן בחרתי בו.

שאלה 7

כיוון שלא ניתן היה להציע שיפורים הכוללים עיבוד מקדים לאלגוריתם (ובפרט נרמול), חשבתי אילו שיפורים יש לאלגוריתם עצמו. השתמשתי בעצי ID3 (מהסעיפים הקודמים) ולא ניתן לשפר את האלגוריתם שלהם כי הוא סגור, ולכן חשבתי מה ייחודי באלגוריתם KNN-Decision-Tree המוצע, אותו ניתן לשפר – אלו הם ה-Centroids ופונקציית המרחק בה משתמשים.

כל שיפור ל-Centroids בסגנון בחירת תכונות הינו עיבוד מקדים בתחפושת (שממנו נמנעתי) ולכן הצעתי את השיפורים הבאים:

1. השיפור הראשון הוא בחירת ממוצע קטום בשיעור (שנע בתחום ) בזמן יצירת ה-Centroids בפונקציית הלמידה (fit). הרעיון היה להתעלם מאחוז קטן מהערכים הקיצוניים הנחשבים חריגים ביחס למדגם (ערכי הדוגמאות האחרות עבור אותה תכונה), אך הצעה זו לא הראתה שיפור.
2. שיפור שני אותו הצעתי הוא משקול התכונות בדמיון על ידי לקיחת בין הדוגמא לסיווג לבין ה-Centroids, כך שעצים קרובים יותר יקבלו משקל גבוה יותר.

**----להוסיף תוצאות----**

1. שיפור נוסף הוא שימוש במרחקים **שאינם אוקלידיים**, כגון: מנהטן; מקדם המתאם של פירסון ו-Jaccard index.

**ניסוי 1: מרחקי מנהטן**

ממוצע 0.9803343136