**חלק א'**

**שאלה 1**

1. נגדיר את ההסתברות להיות -
2. נראה כי מתקיים –

כפי שראינו בתרגול, תוספת האינפורמציה היא תמיד אי שלילית. לכן –

לכן –

בנוסף –

ולכן –

נראה כי מתקיים –

1. כדי להראות שההסתברות עולה כל שתוספת האינפורמציה גדלה, נראה ששיפוע הפונקציה חיובי. לכן נבצע גזירה של הביטוי לפי תוספת האנרגיה -

הנגזרת חיובית ולכן השיפוע חיובי ולכן הפונקציה מונוטונית עולה ביחס ל-.

1. נראה כי לכל תכונה מתקיים :

הראנו כבר שלכל מתקיים . נשאר להראות שמתקיים . נניח בשלילה ש. לכן –

קיבלנו סתירה לכך שתוספת האינפורמציה היא אי שלילית. לכן .

**הרצת ניסוי**

הערכים בהם השתמשנו להרצת הניסוי הם –

גדלי ועדה -

אחוזי סינון -

מספר איברים בעלה -

מספר folds -

רעש –

1. ניתן לראות עבור שני מאגרי המידע, כי הדיוק של הסיווג עולה ככל שגודל הועדות עולה. עם זאת, קצב עליית הדיוק יורד והוא ממשיך לעלות עד לערך רוויה כלשהו. יוצאים מן הכלל הם הסיווגים שהתקבלו על ידי ועדות 1ב' ו-3ב' בסט הפרסומות. עבור ועדות אלו, הערך ההתחלתי הינו גבוה מאוד ולכן קיימת עליה קטנה בלבד.   
   הסיבה שככל שמספר העצים בועדה עולה מקבלים דיוק רב יותר, היא שמספר גדול של מסווגים קיבל את אותה ההחלטה ולכן קיים סיכוי גדול יותר שההחלטה הזאת נכונה. שיטה זו דומה לשיטת ה-boosting בכך שהיא לוקחת אוסף של מסווגים חלשים יחסית ומשתמשת בהם יחד לקבלת סיווג מיטבי.
2. נפריד לשני ה-Datasets –

* **סט הפרסומות** – בסט הפרסומות ניתן לראות שסינון הדוגמאות לא עזר ואף גרע מהדיוק. הדיוק עולה עם עליית גודל הועדה אך מגיע מהר מאוד לרוויה, אותה מקבלים בכל מקרה בפיצול רנדומלי. לדעתנו בסט זה יש תכונות בעייתיות שגורמות לדיוק נמוך, וסינון דוגמאות לא עוזר לפתור את הבעיה הזו, ורק גורם לאימון להיות פחות טוב.

לעומת זאת, סינון התכונות שיפר משמעותית את הדיוק כבר מגודל ועדה קטן מאוד. לדעתנו הסיבה היא שקיימות תכונות בעייתיות, וברגע שמסננים אותם הדיוק עולה כך שועדות 1ב' ו-3ב' קיבלו דיוק גבוה אף מ-ID3. בנוסף ניתן להסביר את התזזיתיות של ועדה 3ב' בכך שהפיצול בה נעשה באופן סמי רנדומלי, ולכן עדיין יש סיכוי שיבחרו תכונות בעייתיות בפיצול הצמתים.

* **סט HAR** – בסט זה ניתן לראות שהחל מגודל ועדה של 5, שיטות הסינון השונות לא משפיעות על הדיוק. כלומר, הדיוק עולה באותה צורה עבור כל סוגי הועדות. ניתן לראות שמתחת לגודל ועדה של 5, יש הבדל בין סוגי הועדות השונות, אך מה שמשפיע הוא לא שיטת הסינון של המידע אלא שיטת הפיצול של הצמתים. כלומר בסט זה אין חשיבות גדולה לסינון המידע.   
  לדעתנו הסיבה להתנהגות זו היא שבסט זה התכונות מתבססות על מידע שמתאר התנהגות פיזיקלית אמיתית שמתארת ישירות את אופן הסיווג (כלומר יש קשר ישיר בין ערכי החיישנים לקביעה האם הבן אדם נמצא בתנועה או לא), כך שקיימות כנראה פחות תכונות בעייתיות. בנוסף בסט זה יש כמות גדולה של דוגמאות (כ-10000), כך שסינון הדוגמאות כדי להתמודד עם הרעש הוא מיותר.

1. נפריד לשני ה-Datasets –

* **סט הפרסומות** – בסט זה ניכר כי ההשפעה המשמעותית היא בשיטת הסינון ולא בשיטת פיצול הצמתים, ככל הנראה כיוון שיש תכונות בעייתיות. עם זאת, ניתן לראות שגם אם מסננים תכונות, כאשר מפצלים צמתים בצורה רנדומלית מקבלים דיוק נמוך. הסיבה היא ככל הנראה שקיימות מספר תכונות קטן שלהן יש משמעות אמיתית לגבי הסיווג, ואם בוחרים תכונות לפיצול בצורה רנדומלית, יש סיכוי נמוך שנקלע לתכונות החשובות באמת. מצד שני בשיטות הפיצול המסתכלות על ה-IG (IG וסמי רנדומלי), אחוזי הדיוק עולים באופן משמעותי.  
  בנוסף, ניתן לראות שאם לא מסננים תכונות, עדיף לבצע בחירה רנדומלית או סמי רנדומלית על פני בחירה לפי IG. לדעתנו הסיבה היא שהתכונות הבעייתיות "מפילות" את ה-IG כיוון שהן מקבלות IG גבוה אך לא באמת מעידות על הסיווג עבור סט הבוחן. לעומת זאת אם בוחרים באופן רנדומלי או סמי רנדומלי, ניתן להתחמק מהתכונות הללו לפחות באופן חלקי.
* **סט HAR** – כפי שראינו בסעיף הקודם, החל מועדות בגודל 5 אין השפעה לשיטת הפיצול. עם זאת, ניתן לראות שעבור ועדות קטנות יותר, עדיף לפצל לפי השיטה הסמי רנדומלית או הרנדומלית על פני פיצול לפי IG. לדעתנו הסיבה לכך היא שלהרעשה יש אפקט משמעותי על ה-IG עבור גדלי ועדה קטנים מאוד, אך האפקט שלה יורד ככל שעולים בגודל הועדה. לכן בחירה סמי רנדומלית מתגברת במידה כלשהי על הרעש הנ"ל על ידי הוספת אקראיות.  
  בשיטה הרנדומלית יש עבור גודל ועדה של 3 דווקא ירידה בדיוק, דבר שיש סיכוי שיקרה כיוון שהבחירה היא אקראית. כאשר גודל הועדה עולה, יש חשיבות נמוכה יותר לכך שהבחירה היא אקראית ולכן רואים עליה דומה לשאר הסוגים.

1. עבור ה-Datasets שקיבלנו, השיטות האופטימליות הן סינון לפי features ופיצול צמתים לפי ה-Information Gain או בשיטה הסמי רנדומלית. בשתי שיטות הפיצול ברור של-IG יש משמעות גדולה לגבי איכות העץ, כיוון שגם השיטה הסמי רנדומלית מתבססת עליו. עבור פיצול סמי רנדומלי, ועבור גדלי ועדה קטנים, מקבלים דיוק די דומה עבור הפרסומות ודיוק גבוה משמעותית עבור HAR. לכן עבור גדלי ועדה קטנים היינו בוחרים בפיצול סמי רנדומלי. לעומת זאת עבור גדלי ועדה גדולים יותר, קיבלנו תוצאות די תזזיתיות עבור הפיצול הסמי רנדומלי בסט הפרסומות, לעומת תוצאות דיוק גבוהות ועקביות עבור הפיצול לפי ה-IG. לכן עבור גדלי ועדה גדולים היינו בוחרים בפיצול לפי IG.  
   לדעתנו הסיבה לכך היא שהוספת אקראיות בגדלי ועדה קטנים עוזרת להתמודד עם רעשים, אך פחות הכרחית עבור גדלי ועדה גדולים שם אנחנו רואים סוג של "חוכמת המונים" שמצליחה להתגבר על הרעש.
2. כפי שראינו בסעיף ד', סינון לפי תכונות ופיצול לפי IG או פיצול סמי רנדומלי הם האלגוריתמים העדיפים מבין השבעה. הפיצול הסמי רנדומלי עדיף תמיד על פני עץ יחיד (ID3) ופיצול לפי IG עדיף כמעט תמיד (פרט לגדלי ועדה קטנים מאוד ב-HAR). החל מגודל ועדה של 1 עדיף להשתמש בפיצול סמי רנדומלי עם סינון לפי תכונות, והחל מגודל ועדה של 5 עדיף להשתמש בפיצול לפי IG וסינון לפי תכונות.  
   בסט HAR, עץ ID3 שקול לועדה בגודל 1 עם פיצול לפי IG, אך לועדה נוספת פעולת הסינון של הדוגמאות. פעולת הסינון של הדוגמאות לפי תכונות או לפי דוגמאות ככל הנראה פוגעת ביכולת הסיווג עבור סט זה. לכן עבור גודל ועדה של 1, אם מבצעים שימוש ב-IG, עדיף לנצל את כל המידע (כל הדוגמאות וכל התכונות) על פני סינון שלו, על אף הרעש.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ad Dataset** | | | | | | | |
| **Acc(ID3)** | **ועדה 3ב'** | **ועדה 2ב'** | **ועדה 1ב'** | **ועדה 3א'** | **ועדה 2א'** | **ועדה 1א'** | **גודל הועדה** |
| 0.932597 | 0.947256 | 0.860018 | 0.962783 | 0.79878 | 0.860976 | 0.744512 | **1** |
| 0.932597 | 0.956098 | 0.860327 | 0.97042 | 0.846951 | 0.862195 | 0.810976 | **3** |
| 0.932597 | 0.930448 | 0.860326 | 0.979872 | 0.858537 | 0.862805 | 0.839024 | **5** |
| 0.932597 | 0.985972 | 0.860021 | 0.976514 | 0.860366 | 0.862805 | 0.84939 | **7** |
| 0.932597 | 0.925303 | 0.860021 | 0.973464 | 0.861585 | 0.862805 | 0.857317 | **9** |
| 0.932597 | 0.93534 | 0.860021 | 0.978646 | 0.862805 | 0.862805 | 0.859756 | **11** |
| 0.932597 | 0.976828 | 0.860021 | 0.971638 | 0.862805 | 0.862805 | 0.862195 | **13** |
| 0.932597 | 0.984723 | 0.860021 | 0.975904 | 0.862805 | 0.862805 | 0.861585 | **15** |
| 0.932597 | 0.973167 | 0.860021 | 0.978652 | 0.862805 | 0.862805 | 0.860976 | **17** |
| 0.932597 | 0.980486 | 0.860021 | 0.979261 | 0.862805 | 0.862805 | 0.862195 | **19** |
|  |  |  |  |  |  |  |  |
| **HAR Dataset** | | | | | | | |
| **Acc(ID3)** | **ועדה 3ב'** | **ועדה 2ב'** | **ועדה 1ב'** | **ועדה 3א'** | **ועדה 2א'** | **ועדה 1א'** | **גודל הועדה** |
| 0.737448 | 0.740046 | 0.742149 | 0.647535 | 0.742028 | 0.729666 | 0.638877 | **1** |
| 0.737448 | 0.806704 | 0.806951 | 0.801383 | 0.814991 | 0.717653 | 0.794215 | **3** |
| 0.737448 | 0.852711 | 0.846279 | 0.857534 | 0.845661 | 0.844796 | 0.848262 | **5** |
| 0.737448 | 0.880536 | 0.870022 | 0.878432 | 0.873857 | 0.8814 | 0.874104 | **7** |
| 0.737448 | 0.892903 | 0.893027 | 0.893275 | 0.889937 | 0.892537 | 0.89847 | **9** |
| 0.737448 | 0.913801 | 0.909973 | 0.915412 | 0.909226 | 0.910588 | 0.910586 | **11** |
| 0.737448 | 0.924313 | 0.923325 | 0.91578 | 0.920232 | 0.915159 | 0.918007 | **13** |
| 0.737448 | 0.928521 | 0.926664 | 0.930003 | 0.929384 | 0.929013 | 0.927037 | **15** |
| 0.737448 | 0.935074 | 0.941134 | 0.933465 | 0.934457 | 0.938908 | 0.933465 | **17** |
| 0.737448 | 0.939403 | 0.943731 | 0.94534 | 0.944102 | 0.943361 | 0.93928 | **19** |

**חלק ב'**

הערכים בהם השתמשנו להרצת הניסוי הם –

גדלי ועדה -

אחוזי סינון -

מספר איברים בעלה -

מספר folds -

רעש –

1. נפריד לשני ה-Datasets –

* **סט HAR** – עבור סט זה רואים באופן ברור עבור כל סוגי הועדות כי ככל שמידת הרעש עולה, כך הדיוק יורד, כאשר הירידה היא לוגריתמית. עבור סוגי הועדות השונים ההתנהגות היא זהה ואף הדיוקים שמקבלים עבור רעשים שונים הוא כמעט זהה. הסיבה לירידה נובעת ככל הנראה מכך שהרעש פוגע ביכולת ההכללה כיוון שהוא מכניס מידע מוטעה שגורם לפיצולים לא נכונים בעצים.
* **סט הפרסומות** – עבור סט זה קיבלנו תוצאות מעט מפתיעות. עבור מרבית סוגי הועדות, הרעש לא השפיע באופן משמעותי על דיוק הסיווג, שהיה נמוך מלכתחילה (בסביבות 86%). עם זאת, עבור סינון התכונות ושימוש בפיצול סמי רנדומלי או לפי IG, קיבלנו דווקא עליה בדיוק ככל שהרעש היה גדול יותר.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ad Dataset** | | | | | | | |
| **Acc(ID3)** | **ועדה 3ב'** | **ועדה 2ב'** | **ועדה 1ב'** | **ועדה 3א'** | **ועדה 2א'** | **ועדה 1א'** | **מידת הרעש** |
| 0.948762 | 0.872525 | 0.860019 | 0.944799 | 0.856098 | 0.856098 | 0.856098 | **0** |
| 0.99878 | 0.893566 | 0.860016 | 0.963405 | 0.861585 | 0.861585 | 0.861585 | **0.05** |
| 0.99878 | 0.921632 | 0.860028 | 0.999695 | 0.859756 | 0.859756 | 0.859756 | **0.1** |
| 0.990854 | 0.98138 | 0.860024 | 0.996951 | 0.851829 | 0.851829 | 0.85122 | **0.15** |
| 0.973773 | 0.980182 | 0.86001 | 0.988718 | 0.85122 | 0.85122 | 0.851829 | **0.2** |
| 0.958221 | 0.963108 | 0.860021 | 0.982618 | 0.853659 | 0.853659 | 0.853659 | **0.25** |
| 0.93595 | 0.984143 | 0.860021 | 0.97621 | 0.862805 | 0.862805 | 0.862195 | **0.3** |
|  |  |  |  |  |  |  |  |
| **HAR Dataset** | | | | | | | |
| **Acc(ID3)** | **ועדה 3ב'** | **ועדה 2ב'** | **ועדה 1ב'** | **ועדה 3א'** | **ועדה 2א'** | **ועדה 1א'** | **מידת הרעש** |
| 0.999011 | 0.999505 | 0.999505 | 0.999505 | 0.999876 | 0.999505 | 0.999753 | **0** |
| 0.949914 | 0.999258 | 0.999382 | 0.999629 | 0.999011 | 0.999258 | 0.999629 | **0.05** |
| 0.907372 | 0.997774 | 0.998268 | 0.998887 | 0.999011 | 0.999134 | 0.999258 | **0.1** |
| 0.863718 | 0.998021 | 0.997279 | 0.997279 | 0.997403 | 0.997774 | 0.997898 | **0.15** |
| 0.815115 | 0.991343 | 0.989364 | 0.990106 | 0.991467 | 0.989736 | 0.989487 | **0.2** |
| 0.77282 | 0.976749 | 0.975143 | 0.979966 | 0.979471 | 0.974771 | 0.975142 | **0.25** |
| 0.722861 | 0.94398 | 0.945214 | 0.942495 | 0.941256 | 0.945831 | 0.948801 | **0.3** |