**עבודה 1 – מבוא ללמידה וניתוח של מידע רב**

שאלה 2:

1. גרפים:
2. ניתן לשים לב בגרף שככל שמדגם האימון גדול יותר הטעות קטנה. הסיבה לירידה בטעות נובעת מהנוסחה הבאה:

גודל ה-H הדלתא והשגיאה המינימלית קבועים וסופיים, כלומר אינם מושפעים מהניסיונות השונים.

ניתן לראות שככל שמגדילים את m מקטינים את הביטוי הכולל של השגיאה.

נשים לב כי הטסט דאטה הוא מדגם מייצג של ההתפלגות, לכן השגיאה עליו מייצגת את השגיאה על ההתפלגות.

1. קיבלנו תוצאות שונות עבור הרצות שונות על אותו גודל מדגם, זאת מכיוון שהמדגם נבחר רנדומלית. בחירה רנדומלית זו גוררת לכך שמדגמים מסוימים יכולים לייצג את ההתפלגות בצורה מדויקת יותר, ולכן עבורם השגיאה תהיה קטנה יותר.
2. ניתן לראות שעבור מדגם למידה קטן ההפרש בין החסם העליון לתחתון שואף לאפס, משום שהמודל כלל לא מייצג את ההתפלגות, ולכן השגיאה גבוהה מאוד בכל הניסיונות.

עבור המדגם הגדול ביותר ניתן לראות שמתרחשת תופעה דומה, ההפרש בין החסמים גם שואף לאפס. זאת משום שבמקרה זה מדגם הלימוד גדול, ולכן כל קבוצה שונה של דוגמאות בגודל זה תייצג את ההתפלגות בצורה טובה מאוד ולכן השגיאה תהיה נמוכה מאוד בכולם.

וניתן לראות שככל שמגדילים את המדגם אז השיפור גדל בהדרגה, אך עבור מדגמים בינונים ניתן לראות פער גדול יחסית בין החסמים. הגורם לכך הוא שעבור מדגם לא גדול, יש המון השפעה לבחירת המדגם. ישנם מדגמים היותר מייצגים את ההתפלגות וישנם מדגמים אשר כמעט ולא מייצגים את ההתפלגות.

1. גרף:
2. השוואה בין גרפים:

שאלה 3:

1. The Distribution of

|  |  |  |
| --- | --- | --- |
| Probability | Genre | Height |
| 33% | Drama | 160 |
| 30% | Comedy | 160 |
| 17% | Drama | 180 |
| 20% | Comedy | 180 |

1. מכיוון שאנו דוגמים רנדומלית את המדגם שלנו, ניתן להשתמש בנוסחה הבאה:

שאלה 4:

שאלה 5:

שאלה 6:

1. לפי ההגדרה של , ההתפלגות היא לא בהכרח realizable לפי . הסיבה לכך היא שיש מקרים בהם היא אגנוסטית, לדוגמא:

התפלגות שבה ישנם שני גרפים בעלי אותה פונקציה g, אך עם תיוג שונה. לא קיימת היפותזה המחלקת ההיפותזות אשר תחזיר לשני הגרפים את התיוג הנכון שלהם. לכל היפותזה שנבחר הם יקבלו את אותו תיוג בדיוק. ולכן לא ניתן לקבל שגיאה אפס על ההתפלגות שתכיל שני גרפים כנ״ל.

עבור התפלגות שבה יש תיוג 1 רק עבור גרפים בעלי אותו ערך g, קיימת היפותזה במחלקת ההיפותזות בעלת שגיאה אפס, ולכן היא ריאליזבל.

1. מימד ה-VC הוא 1. הוכחה:

כאשר גודל הגרף הינו 1, אז קיים גרף יחיד ולכן לכל התפלגות על גרף זה קיימת היפותזה מתאימה במחלקת ההיפותזות.

כאשר גודל הגרף הינו 2 קיימים 2 גרפים אפשריים אשר מקיימים:

בהתפלגויות בהן לשני הגרפים יש את אותו תיוג (1 או 0, לא משנה) אז בהכרח לא קיימת אף היפותזה במחלקת ההיפותזות אשר לא תטעה בכלל (כפי שהסברנו בסעיף א׳).