# דוח הכנה מערכות לומדות 1

נתנאל רוטשילד 204937841

גיא פרג 201096534

1. א. כפי שקראנו במסמך של המעבדה, לא מומלץ לעשות “overfitting”, יש להמנע מבנית מודל שנצמד יותר מידי לדגימות הבוחן שלנו, או לחילופין יכול ליצר בעייה דומה שימוש ביתירות של features שבאופן הגיוני לא יתרמו להבחנה שנרצה שמודל יזהה. לדוגמא מספר טלפון הינו משהו שרירותי ולא יספק לנו מידע על המין של הדגימא. מאפיינים נוספים שנמנע מבחירה בהם הם: צבע שיער וצבע עיניים (ומספרי טלפון).

שאר המאפיינים יכולים לעזור למודל להבחין בין נשים וגברים (גובה, משקל, אורך שיער, ומידת נעליים).

ב. בהינתן סדרת לימוד של דגימות עם תיוגם המתאים נרצה למצוא מסווג בינארי כך שיסווג את סדרת הבוחן למחלקה המתאימה עם שגיאה קטנה קטנה ככל שניתן.

מרחב הקלט:

מרחב הפלט:

1. ה. השיטה tf-idf ככל הנראה שיטה יוצר טובה למציאה של מאפיינים רלוונטים. בחירה בשיטה הזאת תסנן מילות קישור שלא יתרומו הרבה לסיווג המשפט.

*ו. שיטה נוספת לעיבור מקדים יכולה להיות בחירת מילים המופיעות פעם אחד בלבד ובנוסף לא מופיעות בכל המשפטים. בדוגמאות שלנו זה יכול לזהות את שם הדמות במשפט לדוגמא.*

1. *א. נסווג למחלקה שתתן לנו את ההסתברות המקסימלית (סיווג בייסיאני)*

*מכאן אנחנו רואים כי עבור נקבל את הערך המקסימלי ואותו נבחר כסיווג*

*ב. כעת נחשב את הMLE*

*מכאן נקבל כי:*

1. *נראה את זה על ידי שימוש במשערך הסבירות המרבית*

*נמצא את ה*

*כעת נחפש את הערך המקסימלי (נגזור)*

*אפשר לראות כי קבלנו ביטוי שיורד באופן מונוטוני כך שנקבל את הערך המקסימלי עבור ה המינימלי, כאשר החסם שלנו הוא*

1. *נתחיל ראשית בביטוי יחיד*

*או לחילופין:*