בה"ו

טופס הסבר מלווה לפרויקט גמר

הדמיית נתונים סמסטר ב' תשפ"א

מגיש: אורי דרשן

בדף זה נמצאים הסברים לטכניקות ומושגים רבים בהם השתמשתי במהלך העבודה.

סדר ההסברים לפי סדר הופעתם בעבודה (החל ממחברת סיווג, fashion-mnist, כלבים נגד חתולים ומחברת ידיים)

- **Logistic Regression-** האלגוריתם מוצא מודל לוגריתמי (בצורת האות s) בעל סבירות מקסימלית ביחס לנתונים (כאשר מקרה א' הוא המספר 0 ומקרה ב' הוא המספר 1). לאחר בניית המודל, כל מקרה מקבל את הסבירות להימצאותו בכל מקרה לפי המרחק שלו מהמספר שהוצמד למקרה. המקרה בעל הסבירות הגבוהה ביותר הוא המקרה שהמודל ינבא.
- **Voting-** מודל המבוסס על מודלים קודמים, המודל מסווג מופעים באמצעות שילוב של הסיווגים עליו הוא בנוי באמצעות הצבעה. המקרה שסווג הכי הרבה הוא המקרה שהמודל ינבא. ניתן גם להתחשב בסבירות הניחושים של המודלים הפנימיים במהלך ההצבעה.
- **AdaBoost-** מודל בוסטינג- מורכב ממודלים רבים הבנויים אחד על סמך השני. ראשית האלגוריתם מבצע סיווג בעזרת מודל פשוט (לומד חלש), לאחר מכן האלגוריתם מגדיל את המשקל של המופעים שלא סווגו נכון כדי שהמודל הבא יטה לכיוונם. התהליך חוזר על עצמו עד לקבלת אוסף לומדים חלשים. סיווג של מופעים מתבצע בעזרת שקלול של כל הלומדים החלשים.
- **Clustering-** למידת מכונה לא מפוקחת. אלגוריתמים בהם המחשב מחלק את הנתונים למספר קבוצות (מקרים) בהתאם לנתון מוקדם/ בהתאם לנתונים.
- **K-means-** אלגוריתם קלאסטרינג הפועל באמצעות מציאת נקודות המצמצמות למינימום את המרחק הכולל בינן לבין המאורעות הנמצאים בשטחן (השטחים מחולקים לפי המרחק מכל נקודה, כל מאורע משתייך לנקודה הקרובה אליו ביותר).
- **Decision Tree-** אלגוריתם המבצע סיווגים באמצעות רצף החלטות בינאריות המשכיות (ניתן לייצג את תהליך הסיווג באמצעות עץ בינארי), האלגוריתם שואל את השאלות שיצמצמו באופן מירבי את אי הוודאות

של הניחוש הנוכחי. נהוג להגביל את גובה העץ (מספר השאלות שנשאלו) כדי למנוע התאמת יתר למאורעות עליהם המודל נבנה.

- **Random Forest-** אלגוריתם אנסמבל, הבונה מספר לומדים חלשים כדי לבצע החלטות (המודלים לא בנויים אחד בעקבות השני כמו ב AdaBoost), האלגוריתם בונה עצי החלטה רדודים, כל עץ על חלק אחר מתוך כלל המאפיינים של המאורעות. לבסוף האלגוריתם מאחד את העצים לכדי מודל אחד שמשקלל את התוצאות של כל העצים לכדי סיווג בודד.
- **PCA-** אלגוריתם להפחתת ממדים. האלגוריתם מנתח את המאפיינים של המאורעות הנתונים ומוצא/ יוצר מאפיינים המכילים את הכי הרבה מידע (הכי הרבה שינוי ביחס למאורעות השונים). בדרך זו ניתן "לדחוס" מידע רב למספר ממדים מופחת. (מספר ממדים= מאפיינים של המאורעות)
- **XGBoost-** אלגוריתם בוסטינג, בדומה לAdaBoost. האלגוריתם מאמן את המודל הבא שלו על ההפרשים בין המודל הקודם לתוצאות האמת. כל מודל הוא לומד חלש ובסוף האלגוריתם מסווג לפי שקלול של כל הסיווגים שהתקבלו במודלים הפשוטים.
- **Stacking-** בדומה לVoting, גם אלגוריתם זה מיועד לאיחוד של מספר מודלים קודמים. בניגוד להצבעה, כאן ההחלטה הסופית מתקבלת בעזרת מודל נוסף (לרוב רגרסיה לוגית) המקבל את תוצאות המודלים הפנימיים כמאפיינים של מאורעות ומתאמן עליהם.
- **KNN-** אלגוריתם המסווג מאורעות חדשים לפי קרבתם למקרים ידועים. בודקים מהם השכנים של המאורע החדש ומסווגים אותו לפי המקרים הנפוצים ביותר בקרב השכנים (מספר השכנים ניתן לבחירה).
- **Bagging-** אלגוריתם אנסמבל, המאמן מספר מודלים פשוטים על חלקים שונים מכלל המאורעות כדי לקבל מסווגים שונים ולאזן בין נטיות שיכולות להתפתח בכל אחד מהמודלים בפני עצמו. האלגוריתם לעיתים מיישם עצי החלטה כמודלים פנימיים, במקרה כזה האלגוריתם דומה מאוד לRandom Forest, ההבדל הוא ש Bagging מאמן כל מודל על חלק אחר מהמאורעות בעוד ש Random Forest מאמן כל מודל על חלק אחר מהמאפיינים של המאורעות.