

טופס הסבר מטלת גמר סמסטר ב'

תאיר מזריב 209188382

המודלים שהשתמשתי בהם במחברות הם:

Knn:

בהינתן קלט של דוגמה חדשה, האלגוריתם משייכה לקבוצה. הדוגמה משויכת למחלקה הנפוצה ביותר בקרב k השכנים הקרובים) כאשר k מוגדר כמספר חיובי שלם, בדרך כלל מספר קטן. (אם $k=1$ האובייקט משויך למחלקה של השכן הבודד הקרוב ביותר).

Decision Tree:

האלגוריתם הוא עץ בינארי מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ועלים המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם.

Random Forest:

האלגוריתם פועל על ידי בניית ריבוי עצי החלטה בזמן האימון. עבור משימות סיווג, תפוקת היער האקראי היא המעמד שנבחר על ידי רוב העצים.

AdaBoost:

אלגוריתם המורכב ממודלים רבים הבנויים אחד על סמך השני. 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.

XGBoost:

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:

אלגוריתם זה מיועד לאיחוד של מספר מודלים קודמים. ההחלטה הסופית מתקבלת בעזרת מודל נוסף, המקבל את תוצאות המודלים הפנימיים כמאפיינים של מאורעות ומתאמן עליהם.

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.

KMeans:

אלגוריתם הפועל באמצעות מציאת נקודות המצמצמות למינימום את המרחק הכולל בין לבין המאורעות הנמצאים בשטחן.

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.

הסבר כללי על איך פעלתי במחברות: (ב3 האחרונות)

במהלך המחברות קודם כל הפעלתי את המודלים על pca גבוה יחסית (אבל נמוך יחסית למספר העמודות הקודם), חוץ מבמחברת של הידיים ששם הפעלתי ללא pca כי יש שם 107 עמודות.

לאחר מכן בדקתי עם טווח של pca נמוכים יותר על מנת לבדוק אם יש אפשרות להגדיל את היעילות עם הורדה מינימלית של הדיוק.

בסוף בחרתי את המודל שנותן את הדיוק הכי טוב והראתי את הדיוק שלו על הטסט עם pca הנמוך יותר והגבוה יותר. זה תלוי במה מעדיפים: דיוק גבוה יותר ומודל פחות יעיל או דיוק נמוך יותר אך מודל יעיל יותר.