

האקתון IML - 2023

שקד מישורי, לוטם רוזנר, אסף פלדמן ואור מסונייה

במסגרת ההאקתון קיבלנו מידע רפואי אודות ביקורים רפואיים של חולות סרטן השד בישראל. נתבקשנו לנתח מידע זה ולפתח מודל לחיזוי: א) אזורי הגרורות (ב) גודל הגידול. המידע דרש עיבוד מוקדם משמעותי שכן הגיע בצורה מלוכלכת ובעל ערכים שדרשו התאמה והשלמה.

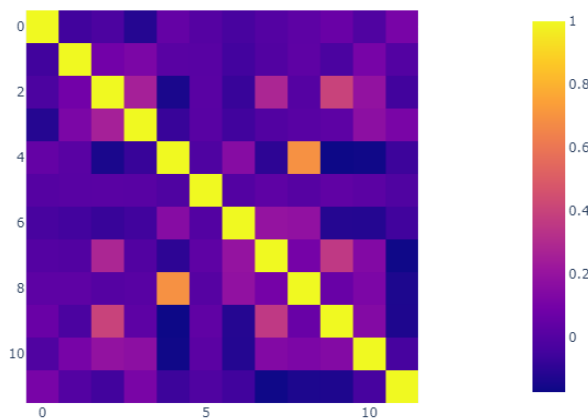
תהליך העיבוד המוקדם וניקוי המידע:

במהלך תהליך העיבוד המוקדם התמודדנו עם מספר בעיות שונות וערכים בעייתיים במידע שדרשו התאמה. השלמת ערכים חסרים – ערכים חסרים של דגימות מסוימות לפיצ'רים או ערכים לא הגיוניים הוחלפו בממוצע עמודת הפיצ'ר.

הורדת עמודות פיצ'רים – עמודות פיצ'רים שרוב ערכיהן היו ריקים הוסרו, כמו גם עמודות לא רלוונטיות לניתוח גודל הגידול או אזורי התפתחות הגרורות (למשל תאריכי ניתוח, שמות בתי חולים, שמות משתמשי רופאים ועוד).

המרת מידע טקסטואלי לנומרי והתאמתו לבעיה – מידע בעל אותה משמעות הגיע בצורה שונה ברשומות רבות ועל כן נדרשנו לסווגם לערכים נומריים אחידים.

בדיקת קורלציה בין פיצ'רים – על מנת לבחון אילו פיצ'רים חשובים יותר לפרדיקציה ואילו פחות הרצנו בדיקה של הקורלציה בין הפיצ'רים השונים והסרנו עמודות שהצביעו על קורלציה שקרובה ל-1.



Hello, we are participating in a Machine Learning Hackathon. Our task is to predict the locations of metastases in breast cancer patients. This classification problem involves multiple labels, with each sample having the potential to be classified into one or more categories. Can you suggest some models that we can utilize for this task?

Certainly! For a multi-label classification problem like predicting the locations of metastases in breast cancer patients, you can consider the following machine learning models:

1. Binary Relevance: This is a simple and straightforward approach where you train a separate binary classifier for each label. Each classifier predicts whether a sample belongs to a particular label or not, independent of the other labels.
2. MultiOutputClassifier: This is a convenient wrapper that can be used with any classifier to extend it for multi-label classification. It treats each label as a separate output variable and trains a separate classifier for each label. It allows you to use any classifier as the base estimator.
3. Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It can be adapted for multi-label classification by using the Binary Relevance or MultiOutputClassifier approach.

משימה 1

בכדי להבין כיצד לבצע משימת קלסיפיקציה מרובת לייבלים

חקרנו את הנושא באינטרנט ונעזרנו בChatGPT:

בחרנו להשתמש במודל MultiOutputClassifier. ניסינו

שימוש במספר לומדי בסיס שלמדנו בקורס כמו עצי

קלסיפיקציה, SVM, KNN ועוד. בחרנו להשתמש בלומד

שהביא ל-F1 score הגבוה ביותר: מודל ה-

RandomForestClassifier. מודל זה משתמש בשיטות

חזקות שלמדנו כמו דגימות Bootstrap ו-Ensemble, ועל כן

הגענו למסקנה שיהיה חזק ורובסטי למשימה.

משימה 2

לאור העובדה שנדרשנו לספק מידע רציף על גודל הגידול השתמשנו במודלי רגרסיה. על מנת לבחור את המודל המתאים ביותר לבעיה, בחנו מספר מודלים והשוונו בין הביצועים שלהם; רגרסיה לינארית פשוטה, רגרסיית ridge, רגרסיית lasso, רגרסיית random forest ומודל פשוט של ממוצע הלייבלים. הרצנו ERM על מודלי הרגולריזציה על מנת למצוא את hyper parameter שממזער את ה MSE וכן הרצנו K-folds cross validation כדי לשפר את ביצועי המודלים. מצאנו כי מודל random forest משיג את ה loss המינימלי על סט המבחן ולכן בחרנו בו לניבוי גודל הגידול.



משימה 3

על מנת לבחון וללמוד דברים חדשים על הדרך שבה המידע מתנהג, בחרנו להריץ מודל Unsupervised Learning של PCA בשני ממדים על מנת לראות כיצד מתנהגים ה principal components. בנוסף בחרנו להריץ מודל של K-means כדי להבין האם המידע שלנו מסווג ל cluster-ים מופרדים. קיבלנו כי הרוב המוחלט של הנקודות מרוכזות לאורך הציר השמאלי האנכי של הגרף, מה שיכול ללמד על כך שנקודות אלו חולקות מאפיינים דומים ובעלי ערכים נמוכים של Principal Component 1. נשים לב גם כי הגרף מציג מספר outliers – נקודות המרוחקות יותר מה cluster המרכזי. נקודות החורגות באופן משמעותי מרוב הנתונים עשויות לדרוש תשומת לב מיוחדת במהלך הניתוח.

