PDF מסכם האקתון IML

לנוכח חתונה, אירועים משפחתיים וימי עבודה של משרות סטודנט היו לכל אחד מארבעתנו שנפלו על תאריכי ההאקתון, התלבטנו מלכתחילה אם נוכל לקחת חלק. קיבלנו החלטה להתאגד יחד ובכל זאת לנסות להפיק מהחוויה הזאת מה שנוכל. לכן ניסינו במסגרת הזמנים המוגבלת מאוד שהייתה לנו ביומיים האחרונים לעשות את הכי טוב שהספקנו, אפילו שנאלצנו להיות במקומות שונים בארץ ולהיות זמינים אחד לשני באופן לא רציף.
נפרט במסמך זה את מה שעשינו וגם את הדברים שחשבנו לעשות והיינו מבצעים אילו היה לנו יותר זמן עבודה משותף:

בהתחשב בכך שהחיזוי שלנו הוא לכל מטופלת, היינו רוצים ששלב ה-pre-processing שלנו יצליח לעשות אינטגרציה בין כל הרשומות (השורות) הרלוונטיות של אותה מטופלת, ולהחזיר רשומה אחת שמכילה את כל המידע הרלוונטי לאותה נבדקת. הדבר נכון גם לסט האימון וגם לסט המבחן שלנו, כך שבסוף התהליך נקבל חיזוי יחיד לכל נבדקת, שאותו היינו מצמידים לכל הרשומות של אותה הנבדקת, ומחזירים את העמודה הזאת בתור הפרדיקציה שלנו. מתוך הבנה שהאילוצים שלנו לא יאפשרו לנו להספיק לעשות את כל התהליכים האלה, החלטנו להשקיע ב-pre-processing כמה שנוכל בשביל שנוכל להריץ עליהם מודל גם אם לא נספיק לבצע את האינטגרציה בין הרשומות של כל מטופלת.
בשלב ה-pre-processing השקענו את מירב הזמן שעמד לרשותנו, והקפדנו לעבוד על פי העקרונות הבאים:

1. בכל עמודה, ראשית ניסינו להבין מה המשמעות שלה ואילו ערכים נצפה לראות בה. אם החלטנו שהיא רלוונטית, פתחנו את הקובץ וראינו בעיניים איך היא נראית - מהם הערכים המצויים בה, כמה מהן מלאות. במידה והיא לא הייתה רלוונטית או שהבנה אם היא אכן רלוונטית דורש ניתוח מעמיק או שינוי משמעותי בצורת ייצוג הדאטה, החלטנו להוריד אותה.
2. בכל אחת מהעמודות האלה, הסתכלנו על התפלגות הערכים של סט האימון במטרה להבין אילו ערכים הם סבירים, מה הם ערכי הקיצון ומה עשויים להיות outliers, וגם להבין אילו ערכים הם בכל הנראה שגויים או עלולים להסיט את המודל שלנו כך שהחיזוי ייפגע.
3. הורדנו עמודות בהן רוב המידע היה חסר ולא היה טריוויאלי מה משמעות תא חסר. כמו כן נאלצנו להסיר גם עמודות שהבנו שיכול להיות להן ערך רלוונטי רב בחיזוי של מיקומי הגרורות או בגדלי הגידולים, אבל הבנו שבמסגרת האילוצים שלנו לא יכולנו להשקיע מספיק זמן בשביל לנתח כמו שצריך את כל חלוקת הערכים האפשריים של העמודה, כמו למשל עמודות ה-markers של החלבונים er ו-pr, בהם היו שלל ערכים אפשריים כמו מספר משתנה של פלוסים ומינוסים, פירוט מילולי של חזק וחלש (בעברית ובאנגלית), באחוזים, במספרים שלמים ובסימני סדר (גדול/קטן).
4. היו עמודות שהכילו מספר גדול של ערכים חסרים, וחיפשנו דרך להשלים את במידע באופן המתאים ביותר. עברנו על העמודות שהכילו ערכי nan, והחלטנו האם מתאים להשלים את ערכים אלו ב-0, ערך הממוצע של העמודה, החציון וכדומה.
5. בעמודות שבהן היו ערכים מילוליים עם יחס סדר יחסית ברור ביניהם, כמו שלבים שונים של ה-stage-אבחנה, סידרנו אותן על רצף מספרי ונתנו להן ערכים מספריים שמשקפים את הסדר.
6. עבור עמודת ה-side-אבחנה, אחרי התלבטות ורצון להימנע מהרבה מאוד דאטא sparse-י, החלטנו להימנע מעמודות dummies ולתת ערך מספרי שמייצג את מספר הצדדים שנכונים לאבחנה - כלומר אם מדובר בצד ימין או שמאל נתנו 1, ואם מדובר בשניהם נתנו 2. אחרת נתנו 0. אנחנו מבינים שבזאת יש החסרה שיכולה להיות חשובה של הדאטא, ככל הנראה בעיקר עבור המשימה הראשונה שרלוונטית לאזורים אפשריים לגרורות, ולכן אנחנו מבינים כי בהינתן יותר זמן ופחות אילוצים היינו רוצים להתייעץ גם כאן עם איש מקצוע רפואי בשביל להבין עד כמה יש רלוונטיות לצד עבור כל אחד מהחלקים של המשימה.
7. בעמודת ה-Nodes exam שמציינת את מספר בלוטות הלימפה שנבדקו, החלטנו להחליף את הערכים בעמודה ביחס שבין מספר בלוטות הלימפה שנמצאו חיוביות לבין מספר בלוטות הלימפה שנבדקו. עשינו זאת מתוך חשיבה שהיחס בין מספר הבלוטות שנבדקו ומספר הבלוטות החיוביות מעיד על החמרה אפשרית של הסרטן ובכל מקרה מעניק לנו יותר מידע מאשר הדאטא המקורי בעמודה.
8. תהליכי pre-processing נוספים שחשבנו לעשות ונאלצנו לוותר עליהם היו השוואת התאריכים השונים והוספת עמודות שמביעות פערי זמנים בין תורים ופגישות, או עמודה שמביעה משקל שונה לסוג הפגישה או טיפול - כאשר תהליכים אלו היו מחייבים אותנו להבין יותר לעומק את עולם

התוכן ולהבין את המשמעות והחשיבות של תורים/טיפולים מסוגים שונים והקשר שלהם לשלב שבו נמצאת המטופלת.

9. כמו כן, בעמודות שבהן יש תיאור מילולי של סוג הטיפול או הניתוח, היינו רוצים להצליח לאתר מילות מפתח ולהכניס אותן כעמודות אינדיקטור, שלדעתנו מניתוח המידע הראשוני שעשינו, היו יכולות לעזור בעיקר בחלק הראשון של חיזוי אזורים אפשריים שבהם יהיו גרורות (כמו למשל איזכור מפורש של בלוטות הלימפה). כמובן שאם היינו נוקטים בצעד זה היינו מחויבים ללמוד יותר על המשמעות של הופעת המילים האלה בעמודות האלה, עד כמה הן טריוויאליות ומה ההשלכות שלהן מבחינה רפואית.

עבור המשימה הראשונה של חיזוי מיקומי הגרורות, החלטנו ללכת באסטרטגיה של לחזות בנפרד לכל מיקום אפשרי רלוונטי את הסבירות שלמטופלת תהיה בו גרורות. לאחר שלב ה-pre-processing וקבלת הדאטא המעובד שלנו בתור סט האימון, בכל פעם נאמן בעזרתו מופע חדש של המודל שלנו עם לייבלים בוליאניים של מיקום אפשרי אחר גרורות. כך למשל נאמן מודל אחד עם לייבלים של 1 אם למטופלת יש גרורות בעור ו-0 אחרת, מודל נוסף עם לייבלים שהם אינדיקטור לגרורות במוח, וכך הלאה... לכל אחד מהמודלים האלה נחזה בנפרד על סט המבחן שלנו ונקבל לכל נבדקת ערך כלשהו בין 0 ל-1 שמעיד באופן כלשהו על הסבירות (לפי המודל) שיש לה גרורות באזור הזה. נבחר להשתמש במודל שמבצע רגרסיה, וכחלק מתהליך הלמידה נצטרך לבחור עבור כל אחד מהמודלים הנ"ל את ה-threshold המתאים שלבסוף יאפשר לנו לקבל החלטת קלאסיפיקציה - האם אנחנו חוזים שיש לה גרורות באיזור הזה או לא (כמובן שה-threshold יהיה שונה לכל מודל). לסיום נאחד בין כל החיזויים שעשינו לכל אחד מהמיקומים, ונחזיר את רשימת המיקומים הרלוונטיים שחזינו לכל נבדקת בתור החיזוי הסופי שלנו לרשומות שרלוונטיות אליה.

עבור המשימה השנייה של חיזוי גודל הגידול, היינו רוצים לבחור לאמן את המודל שלנו רק על רשומות שגדלי הגידולים שלהן הוא בטווח הנפוץ, בשביל לא להסיט אותו מדי לערכי קיצון. נשים לב מניתוח סטטיסטי של הערכים האפשריים מסט האימון, כי מתוך 49,351 הרשומות, 28,568 הן עם גודל 0, כלומר ללא גידול - יותר מחצי מהרשומות של המטופלות. מתוך השאר, כלומר מתוך ה-20,783 עם גידול כלשהו, 10,409 (בערך חצי) הן עם גידול של 0.03 עד 1.7 כולל, כאשר רוב ההתפלגות היא על ערכים שהם בין 1 ל-1.7. יש רק 7 רשומות (שתי מטופלות) עם גודל שקטן מ-0.1, לכן ככה"נ מדובר בגדלי outliers שעדיף לא להסיט בעזרתן את המודל ולא להתאמן עליהן.
יש עוד 8,117 רשומות עם גדלים שהם בין 1.8 ל-3.9, ועוד 2,041 עם גדלים שהם בין 4 ל-8. יש רק 201 רשומות עם גדלים שהם בין 8.1 ל-38, ורק עוד 15 שהם 100.
לכן, אם היה לנו זמן להצליח לממש את כל מה שחשבנו עליו, היינו בוחרים להתאמן רק על הרשומות שהן החל מ-0.1 (כולל) עד ל-4, או עד ל-8, תלוי לפי סוג הטווח שהיה נותן לנו מודל שחוזה טוב יותר כשנבחן אותו על סט ה-validation שלנו.
עבור המודל עצמו, בחרנו להשתמש ברגולריזציה מסוג רידג' עם lambda=0.05. הלמדא נבחרה על פי העקרונות שלמדנו - חילקנו את הסט האימון ל-train, dev, test, ונבחנו מודלים שונים עם פרמטרים שונים באמצעות k-fold. ניסינו Lasso, Ridge, RandomForest, Adaboost ולבסוף, בהתאם ל-loss, בחרנו את ridge עם הלמדא הרצויה - על פי כללי האצבע שנלמדו בקורס, בחרנו בlambda הנמוך שהחל להתייצב ב-loss על ה-dev set.

כעבודת המשך, היינו רוצים להמשיך ולהעמיק בשלב ה-preprocessing ולנצל את המידע שקיבלנו - היינו רוצים לנסות לקבץ את המטופלות לפי המזהה הייחודי שלהן, ולנסות חיזויים שונים בהתאם לייצוגים שונים של תאריכי הבדיקות. בנוסף, לדעתנו יעזור מאוד להתייעץ עם מומחה בתחום סרטן השד כדי להבין ממנו לעומק מה כל פיצ'ר מייצג ומה לדעתו בעל משקל רב יותר בחיזוי סרטן מנסיונו האישי.

מה היינו עושים אם היה לנו זמן:

1. מקבצים לפי מטופלת
2. עבור שאלה 1, יצרנו מודלים שונים עבור כל סוג סרטן, אך בסופו של דבר הרצנו את אותו אלגוריתם עבור כל המודלים, כלומר השתמשנו ב-AdaBoost עבור כל סוגי הסרטן. אם לא היינו מוגבלים בזמן, היינו עושים שימוש באלגוריתמים ומודלים מסוגים שונים ובכך מקטינים את ה-Loss כמה שיותר עבור כל אחד מסוגי הסרטן בנפרד.