Roi Fogler 302882527
Ori Fogler 318732484

2 Dataset

Your data set is about Abalone age classification. In this dataset, you are provided with eight features per instance (seven of which are numerical and one is the categorial) and three labels. Labels correspond to the Abalones age. You should explore different ways to convert this categorial

attribute to a numerical one. Moreover, try different normalization techniques as well as feature selection.

A summarize of the feature description can be found in the table below:

Name	Data Type	Meas.	Description
Sex	nominal	-	M, F, and I (infant)
Length	continuous	mm	longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried

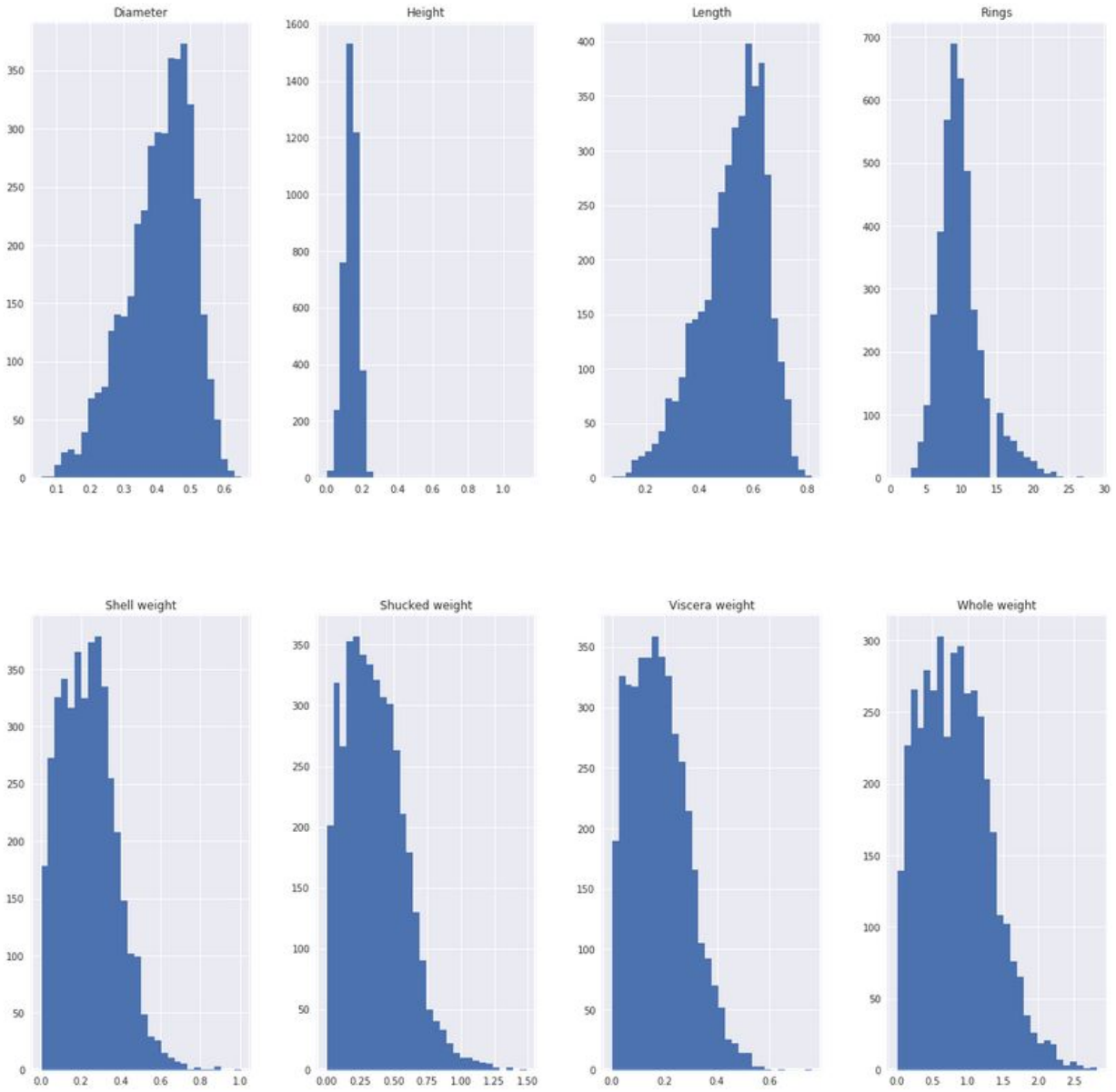
- A PDF report with all the implementation details and hyper-parameters, (How did you choose the learning rate? What values did you try and how was the performance? How did you choose the lambda value for SVM? etc.).

להלן נשתדל לסקור מרכיבים מהותיים בתהליך הלמידה, גם שלנו וגם של התוכנית.
בין השאר, נציג עובדות וניתוחים שקשורים לאופי המידע ותכונותיו. נדון בהשלכות אפשריות של זה.
כמו כן, נציין תוספות וטכניקות שבעזרתן ניסינו לשפר את תהליך הלמידה. למשל ישנן שיטות נרמול שונות, שיטות לוולידציה, מימושים שונים לאלגוריתמים ואפשרות הוספת 1 לווקטור הפיצ'רים בתור BIAS.
נציג גם שיקולים הקשורים לקביעת ערכי ההיפר פרמטרים, ואת ההחלטות שלנו לבסוף לגביהם.

Roi Fogler 302882527
Ori Fogler 318732484

ניתוח הדאטה

קיבלנו מידע שנאסף עבור אלפים בודדים של צדפות.
בין הפיצ'רים ישנו פיצ'ר אחד של "מין". ישנן שלוש אופציות - 'M', 'F', 'I'.
שאר הפיצ'רים נחלקים, בגסות, לשניים - פיצ'רים הקשורים להיקפים ונמדדים במ"מ, ופיצ'רים הקשורים למשקל ונמדדים בגרמים.
נתבונן בדיאגרמות הבאות, ונדון במה שרואים בהן:

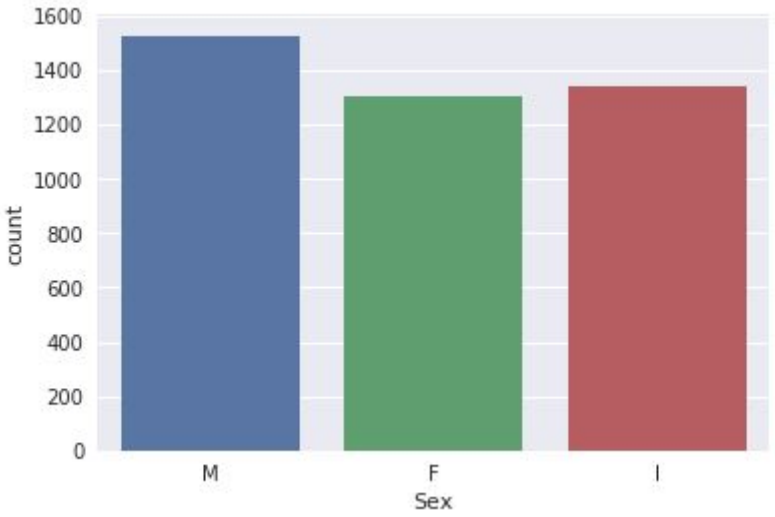


קודם כל הבחנו בכך שישנם ערכים מאופסים בכמה פיצ'רים עבור כמה דגימות.
לא סביר שגובה, למשל, של צדפה יהיה 0, ובטח אם חלק שאר הערכים רחוקים יחסית מ0.
הסקנו שמדובר בטעות. דגימה לא טובה.
עוד הסתכלנו על הטבעות. זהו הפיצ'ר שאצלנו מתחלק לשלוש אופציות (למעשה אצלנו מדובר בתיוגים) - 0,1,2.

Roi Fogler 302882527
Ori Fogler 318732484

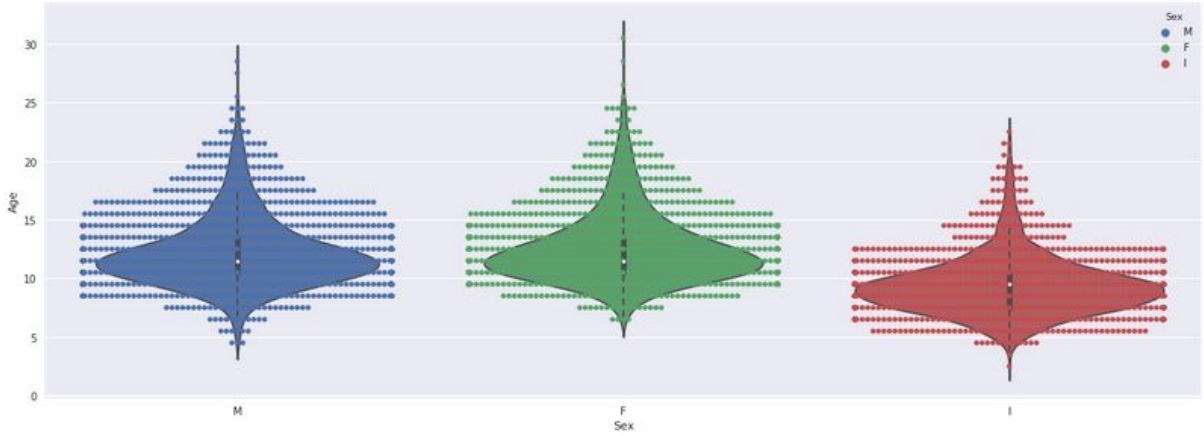
מעניין היה לראות כי אצל רוב הצדפות יש בין 5 ל15 טבעות. מתחת ל5 ומעל ל15יש הרבה פחות, משמעותית.

מה החלוקה של הצדפות למינים? התבוננו בדיאגרמה:



באופן גס, ניתן לומר שהחלוקה בין המינים, פחות או יותר, שווה.

מיד שאלנו את עצמנו, האם יש בכלל הבדל בין המינים (אנחנו פמיניסטים), ואם יש, האם הוא משמעותי? באשר להבדל בין MALES לFEMALES התוצאות היו מאוד מעניינות, כפי המופיע בדיאגרמה:



רואים שהתפלגות הגיל, אצל הזכרים והנקבות, כמעט זהה.
ברגע זה עלו לנו שאלות רבות, כגון - האם בכלל החלוקה בין זכרים לנקבות רלוונטית? אולי מי שחילק את המידע היה צריך לסווג לקטגוריות אחרות?

עוד ראינו, באופן לא מפתיע, שהINFANTS באמת בגילאים הנמוכים יותר.
למעשה, עד גיל 5, כמעט כל הצדפות הן INFANTS.

Roi Fogler 302882527
Ori Fogler 318732484

מה חשבנו לעשות בעקבות ניתוח הנתונים?

עלו לנו רעיונות שונים, למשל לשלול באופן מפורש מINFANTS מלהיות במחלקת גיל גבוהה.
עוד רעיון שעלה, הוא לבטל בכלל את ההבדל בין זכרים לנקבות.
בסמוך נציג את הבעיה של פיצ'ר של קטגוריות לא נומריות ופתרונות שלה, אך אם בכלל לא היינו מחלקים בין זכרים לנקבות, ע"ס הנתונים, כלל לא היו לנו שלוש קטגוריות. היו רק 2, והפתרון היה טריויאלי - פיצ'ר בינארי יחיד של "האם INFANT".

חוץ מזה, כמו שהזכרנו בהתחלה, ישנן דגימות שחשדנו שהן פגומות (אין צדפה בגובה 0), וחשבנו אולי להתעלם מהן באימון.

לבסוף לא הכנסנו את ההצעות לעיל לקוד שלנו, מכיוון שניסינו אותן ולא ראינו תוצאות מובהקות לטובה.
עם זאת, שמחנו על כך שצברנו קצת נסיון וחוויה בלמידה מעמיקה של הנתונים עליהם אנחנו עובדים.

קטגוריות לא נומריות

בעיה שנתקלנו בה בתרגיל זה, היא התמודדות עם פיצ'ר של קטגוריות לא נומריות.
המינים הם זכר, נקבה ותינוק. לא מספרים. האלגוריתמים שלנו, כמובן, דורשים מספרים.
פתרון מיידי לבעיה היה לקבוע שרירותית מספרים לקטגוריות. למשל לקבוע שזכר יהיה 0, נקבה 1 ותינוק 2.
אמנם, יש בעיה בפתרון זה. בין 0, 1 ו2 מתקיים יחס סדר. 2 גדול מ1 שגדול מ0. האם תינוק גדול מנקבה שגדולה מזכר? כמובן שלא. האם הממוצע של תינוק וזכר זה נקבה כמו שהממוצע של 2 ו0 זה 1? לא.

מצאנו פתרון אחר, הנקרא ONE HOT ENCODING. למעשה מדובר בפתרון די פשוט.
במקום פיצ'ר יחיד של קטגוריות מין, הוספנו לדגימות שלושה פיצ'רים, עם ערכים בינאריים - "האם זכר?", "האם נקבה?" ו"האם תינוק?".
נשארנו עם הפתרון הזה, והתרשמנו שהתוצאות עימו טובות יותר.

נרמול

למדנו בתרגול על כך שנרמול עשוי לשפר את התוצאות.
שתי השיטות שלמדנו הן MIN MAX, וZSCORE.
ניסינו שלוש אפשרויות - בלי נרמול, עם נרמול MIN MAX ועם נרמול ZSCORE.
ראינו תוצאות יותר טובות עבור האפשרות השלישית, ולכן בחרנו בה.

שמנו גם לב שצריך לבדוק במהלך הנרמול שלא מתבצעת חלוקה ב0 (או לחלופין לשים TRY ולקחת בחשבון אופציה כזו).
למה שתתבצע חלוקה ב0? למשל אם בפיצ'ר מסויים, עבור כל הדגימות, הערך אותו הדבר. במקרה זה ההפרש בין MIN וMAX יהיה 0. סטיית התקן תהיה גם כן 0.

נתקלנו באינטרנט בתגובה הבאה:

Roi Fogler 302882527
Ori Fogler 318732484
Yes you need to apply normalisation to test data, if your algorithm works with or needs normalised training data*.

That is because your model works on the representation given by its input vectors. The scale of those numbers is part of the representation. This is a bit like converting between feet and metres . . . a model or formula would work with just one type of unit normally.

Not only do you need normalisation, but you should apply the exact same scaling as for your training data. That means storing the scale and offset used with your training data, and using that again. A common beginner mistake is to separately normalise your train and test data.

כמו שכתוב בה, נרמלנו גם את הערכים עליהם אנחנו מתאמנים, וגם את הערכים של הטסט.
בנוסף, כמו שכתוב בפיסקה השלישית, הנרמול צריך להעשות על פי הפרמטרים של האימון. גם הנרמול של הטסט. הקפדנו על כך, ושמענו אחרים שעשו טעות זו ותיקנו אותם.

ביאס BIAS

במקום WX+B, אפשר כמובן להוסיף לפיצ'רים עוד עמודה נוספת עם 1. זה כמובן שקול.
כך, כפי שלמדנו בהרצאה, המפריד הלינארי יכול לעבור לעבור שלא בראשית, וכך תוצאותיו עשויות להשתפר.
ניסינו עם וניסינו בלי, והחלטנו להוסיף זאת.

פרספטרון

בכל האלגוריתמים, כתבנו את המימוש עבור בעיית MULTI CLASS, שכן זו הבעיה בתרגיל הנוכחי.
הפלט אינו מפריד בין 1 ל1-, אלא מסווג לאחת משלוש אופציות - 0,1,2.

תוך כדי העבודה, ניתקלנו במאמר הבא של פרופ' קשת
http://u.cs.biu.ac.il/~jkeshet/teaching/aml2016/multiclass.pdf.
מופיעות שם שתי אופציות עבור פרספטרון.
לא רק עדכון הW עבור התיוג הנכון והתיוג הלא נכון שהחזיר המסווג, אלא עדכון עבור כל התיוגים הלא נכונים.
ניסינו את שתי האפשרויות, והתרשמנו שבאפשרות השנייה הדיוק נמוך יותר, אך ההתכנסות מהירה יותר.

וולידציה

אחרי שאימנו על הדאטה, רצינו לבדוק אם האימון היה בסדר.
יתר על כן, צינו לכייל את ערכי ההיפר פרמטרים להיות הערכים שנותנים את התוצאות המיטביות.

בשלב זה, שאלנו את עצמנו את השאלה בעזרת איזה דאטה עושים את הוולידציה?
אפשרות פשוטה היא לבדוק את האימון על הצדפות שבעזרתן עשינו את האימון.

Roi Fogler 302882527
Ori Fogler 318732484
זה לא אפשרות בלתי סבירה לחלוטין, ואכן ניתן לסנן בעזרתה תוצאות גרועות במיוחד, למשל, אך זו גם, כמובן, לא האפשרות המושלמת. הדבר דומה לאדם שבודק אם הוא עקום, מזווית הראייה של עצמו (או בעברית - OVERFITTING).

ניסינו להפריש חלק מהדאטה שקיבלנו לטובת הוולידציה.
ניסינו חלוקות שונות, למשל 70 אחוז לאימון ו30 לוולידציה, או 90 אחוז לאימון לאימון ו10 לוולידציה.
קשה לומר מה החלוקה האופטימלית של הדוגמאות לאימון ולוולידציה, אבל כך או כך, "קיבלנו כיוון" אם אנחנו משתפרים או נעשים פחות טובים.

אפשרות נוספת, אותה למדנו בתרגול האחרון, היא K CROSS VALIDATION.
לפי שיטה זו, חילקנו את הדאטה ל6 חלקים (למדנו שהנסיון מלמד שזה סדר הגודל הטוב לחלוקה), וכל פעם בחרנו חלק לוולידציה, אימנו על פי החלקים האחרים ובחנו אותם מולו.
לבסוף חישבנו כמובן את ממוצע הדיוק עבור ששת האימונים.
עושה רושם שזו הדרך היותר מדוייקת שבה בחנו את האימון שלנו ואת ההיפר פרמטרים שלו.

לבסוף ערכי ההיפר פרמטרים שבחרנו הם - מספר איטרציות יחסית נמוך. בין איטרציה 1 ל10, באלגוריתמים השונים.
מדובר באלפי דגימות, ולכן גם עשרה מעברים על כל הדגימות זה מספר די גבוה.
100 איטרציות או 500 (כמו ששמענו שאחרים עשו), זה סדר גודל די מוגזם. לא היינו צריכים מספר כ"כ גדול, כיון שראינו התכנסות בשלב הרבה יותר מוקדם.

לגבי ערכי ההיפר פרמטרים האחרים, בהכללה גסה, לא ראינו שינוי משמעותי דרסטי.
כן החלטנו מעט לעדן את הדיוק עם מספר קצת יותר גדול של איטרציות, ועם ערכי קבוע למידה קטנים יותר. למשל 0.001 עבור האתא.
אכן, קבוע רגולציה עם ערך גבוה מדי, הוריד את הדיוק. בחרנו עבורו ערך נמוך יחסית - 0.001.

Roi Fogler 302882527
Ori Fogler 318732484