Markov Networks as Prior Distribution for Multi-Label Classification (in Wikipedia)

#### By: Alon Wolf (318336930) and Aviv Hadar (208948885)

# תקציר - Abstract

נתקלנו בבעיית Multi Label, עבורה יצרנו מודל אשר מפריד כל תיוג בפי עצמו, אליו הוספנו שימוש ב-prior ע"י מודל מרקובי על המחלקות מבוסס אלגוריתם צ'או ליאו. רק שה-prior משפר את תוצאות המודלים, בפרט ובעיקר במודל מבוסס LightGBM. ניסינו ואריציות שונות של אלגוריתם צ'או-ליאו והשוונו אותם עבור מודל עם עץ רנדומלי. ראינו כי שימוש ב-M.I. נותן את התוצאות הטובות ביותר. לבסוף ניסינו להרכיב בין המודלים השונים בשביל להיעזר בשונות בין המודלים לצורך שיפור ה-prior, אמנם ניסיון זה העלה חרס.

# רקע תאורטי

* מודל מרקובי pairwise *// אביב*
* אלגוריתם צ'או ליו *// אביב*
* שימוש אלגוריתם צ'או ליו כ-prior *// אביב*
* "הכללה" לקורלציות שליליות *// אביב*

כפי שציינו קודם, במאמר נעסוק בשיפור ביצועי אלגוריתם בבעיית multilabel בעזרת יצירת prior על ה-labels בעזרת אלגוריתם "Chow Liu" ומודל מרקובי מעל להתפלגות הלייבילים.

לכל דאטאסט multilabel עם לייבלים:

עבור מודל הסתברותי - Pairwise Markov Network על הלייבלים שלו, מתקיים שהנוסחה ל-Inference זו:

ואם נניח שמבנה הרשת הוא עץ נקבל שהמשקלים האופטימליים לרשת הם:

כאשר מתקיים:

לכל classifier נזכור שמעניין אותנו חישוב ה-Marginals כדי לדעת את הסבירות של כל לייבל להיות חיובי. ניתן לחשב Marginals של מודל מרקובי בעזרת שיטת ה-Belief Propagation בצורה הבאה:

כאשר מתקיים:

משוואה (3) בעצם נותנת לנו את ההסתברות לקיומו של לייבל מסוים לפי **התפלגות הלייבלים בלבד**, זאת מפני שהמודל הנ"ל ממשקל את כל ערכי באופן זהה ו- מושפע מ- בלבד (בהינתן הצבת המשקלים האופטימליים). כלומר המודל מתעלם כרגע מהסבירות לערכי לייבל שונים בהינתן מאפייני הדוגמא, שזו יכולה באופן משמעותי להשתנות בין דוגמאות שונות בהתאם לערכי הפיצ'רים.

המטרה שלנו בפרויקט הייתה לבצע fine tuning לביצועי מודל ע"י prior שמשתמש בקונספט של רשתות מרקוביות כדי למדל את התפלגות הלייבילים, הוספנו לנוסחת חישוב ה-Marginal משקול ערכי בכל הודעה לפי הפרדיקציות של המודל - .

כאשר מתקיים:

כלומר לשלב את הסבירות לקיומו של לייבל בהינתן הפיצ'רים () עם הסבירות שנובעת מהתפלגות הלייבלים (), במטרה לשפר את ביצועי המודל.

*הנוסחה המשולבת היא להלן:*

*חשוב לציין ש- אינה פונקציית הסתברות שכן נדרש לנרמל אותה:*

*זו הפרדיקציה לכל לייבל שמוחזרת מהפריור.*

אינטואיציה – חשבתי להוסיף איזה משפט שניים על אינטואיציה ללמה זה הגיוני.

*אלגוריתם צ'או ליאו:*

*באופן הבא:*

1. *הגדרנו גרף קשרים בין הלייבלים עם המשקלים הבאים:*
2. *לאחר מכן מצאנו עץ פורש מקסימאלי לגרף*
3. *והעץ שקיבלנו שימש אותנו לבניית המודל המרקובי*

## Partial MI Chow Liu



*איור 1 – ערך ויקיפדיה של אלטון ג'ון. באדום – הערך (טקסט), בכחול – עובדות מהירות. מטרת הפרויקט היא לחזות את החלק הכחול בהינתן האדום.*

# מידע (דאטה)

את המידע חילצנו מאתר ויקיפדיה. השתמשנו בספריות wikipediaapi, wptools על מנת לחלץ את הטקסט וה-infobox של ערכי הויקיפדיה (דוגמה לערך ניתן לראות ב*איור 1*).

נשאלת השאלה אילו ערכים לקחת – לא היינו לקחת את כל ערכי ויקיפדיה כיוון שמספרם גדול מדי ודגימה רנדומלית מתוך ויקיפדיה אינה ממשית (וכן צריך דגימה גדולה מאוד בשביל שתספיק תייצג את כלל הערכים). על כן, דגמנו את הערכים מ-subset של ערכי ויקיפדיה והוא "מוזיקאים"[[1]](#footnote-1) (בכדי שהערכים ייבחרו על ידי התפלגות יחידה ומוגדרת היטב).

עבור כל ערך בהתפלגות זאת, לקחנו את הטקסט כמשתנה הבלתי-תלוי ואת השאלות של ה-infobox כמשתנה התלוי.

את הטקסט העברנו עיבוד פר מילה כך שלכל מילה ייצרנו ממד חדש על האם הטקסט הכיל את המילה. דוגמא לתהליך זה עבור המילה *'Birth\_date'* ניתן לראות ב*איור 2*. לכל מילה הכנסנו ציון לפי אלגוריתם TF-IDF כך שככל שמילה מופיעה יותר בטקסט כך היא מקבלת ציון גבוה יותר, מנורמל לפי מס' הפעמים שהמילה הופיעה בכלל הטקסטים. הורדנו מילים שחזרו במס' בודד של טקסטים כיוון שלא רצינו שהמודל ילמד קורלציה שרק קורלטיביות למילים בודדות.

*איור 2 – עיבוד המילה על מילה "Birth\_date". ניתן לראות שכל חלק ...*

Tokanization

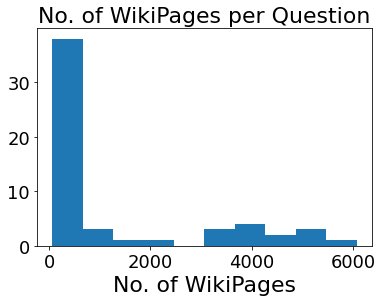
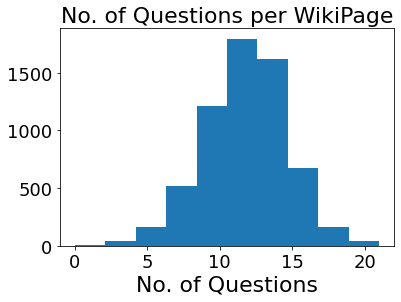
*'Birth\_date'*

Stop Words Deletion

*'Birthdate'*

Stemming

*'Birthdat'*



*איור 3 – היסטוגרמות של תיוגים ומחלקות. כמה שאלות (תיוגים) יש לכל דוגמא וכמה דוגמאות יש לכל שאלה (מחלקה)*

הסיבה שהשתמשנו בעיבוד זה היא כיוון שבשאלות ובטקסט היו עיוותים של מילים שמתארות את אותו הדבר - כן לדוג' *'Birth\_date', 'birth\_date', 'birthdate'* היו מילים שונות בטקסט שלאחר העיבוד שלנו הפכו למילה אחת *'birthdat'*.

כמו כן, מתוך השאלות גם כן סיננו שאלות לא נפוצות (מופיעות עבור מס' בודד של ערכים) שמספר מופען קטן ולכן אינן מייצגות את התפלגות המחלקה.

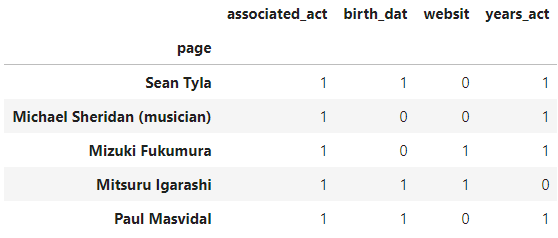
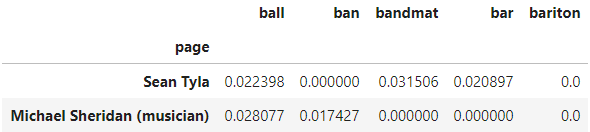
כפי שניתן לראות מהדוג', הבעיה הינה MultiLabel כיוון שלכל ערך ישנו מס' משתנה של שאלות.

את התפלגות מס' השאלות לכל ערך ניתן לראות באיור מס' 3. כפי שניתן לראות מהאיור, הופעת התיוגים מגוונת ואיננה טריוואלית – כלומר יש שאלות שמופיעות עבור עשרות דפים (כגון *'----'*) יש שאלות שמופיעות עבור מאות דפים (כגון *'----'*), ויש שאלות שמופיעות עבור רוב הדפים (כגון *'name'*).

## סיכום המידע

סה"כ קיבלנו 6221 דוגמאות עם 4328 פיצ'רים ו-56 סוגי מחלקות.

דוגמא למידע שקיבלנו ניתן לראות ב*טבלה 1*. ניתן לראות דוגמאות לפיצ'רים שקיבלנו וכן דוגמאות לשאלות עבור 2 ערכים – 'Sean Tyla'[[2]](#footnote-2), 'Michael Sheridan'.



*טבלה 1 – דוגמא למידע שקיבלנו עבור 2 דפי ויקיפדיה. באיור העליון דוג' לשאלות עבור הערכים. באיור התחתון דוג' לפיצ'רים עבור הערכים.*

# תהליך

התחלנו בלייצר מודל לכל מחלקה (שאלה). השתמשנו במודל OneVsRest – כלומר, יצרנו מודל לכל מחלקה בנפרד. עשינו זאת כיוון שזהו common practice שמשתמשים בו כשרוצים למדל בעיות של multilabel. נשים לב שהמודל לא מתחשב בקשרים בין המחלקות כיוון שהוא מנתח כל מחלקה בנפרד.

~~על כן, רצינו להוסיף "prior" אשר כן יתחשב בקשרים בין המחלקות כך שאם יש קורלציה (MI) בין 2 מחלקות אז זה ישפיע על הציונים שהוא נותן. רצינו לדעת הם וכמה אנחנו יכולים לשפר את המודל הראשוני תוך שימוש ב"prior" על הדוגמאות. את ה-prior מימשנו באמצעות מודל מרקובי בין המחלקות.~~

את ה-prior () חישבנו כפי שתיארנו בחלק התאורטי (משוואה (X)).

בשלבים הבאים:

1. את ה- חישבנו לפי סט האימון בלבד בשביל לא לייצר Target Leakage בבחינה של המודל.
2. *בחרנו את מבנה העץ של המודל המרקובי ע"י אלגוריתם Chow Liu.*
3. השתמשנו בעץ וב- בשביל לחשב את על פרדיקציות האלגוריתם ע"י ...

*~~כאמור מהחלק התאורטי, השתמשנו ב-Belief Propagation על מנת לייצר ...~~*

*בהמשך בחנו את המודל תחת מבני עץ שנבנו בצורות שונות מזו, יוצג בחלק הבא.*

*כעת נתאר כיצד בחנו את השפעת המודל הנ"ל על ביצועי אלגוריתם ML.*

השתמשנו במודל סטנדרטי \*ואיזנו את התוצאות\*.

\* יש לשים לב כי לפי הנוסחא ב-, משקול ההודעות החוזרות מתבצע לפי פרדיקציות המודל ו**מניח כי פרדיקציות אלו הן אכן הסתברויות** ולכן, בכדי שייצגו פונקציות הסתברות, אנחנו ממליצים לנרמל את משקלי המחלקות באימון המודל.

# תוצאות

## אופי בחינת מודל

השוואנו את תוצאות המודלים לפי 2 מדדים: , . החלטנו על מדדים אילו כיוון ש-roc\_auc הינו מדד אינטואיטיבי אשר מודד מה הסיכוי שהמודל ייתן לתיוג חיובי ציון גבוה מתיוג שלילי. אמנם, roc\_auc מקבל ציונים גבוהים עבור תיוגים לא מאוזנים וכיוון שיש לנו מס' שאלות עם מעט תיוגים, הוספנו את המדד *כיוון שאינו מתחשב ב- ולכן מייצג בצורה שונה את טיב המודל.*

*השתמשנו ב-2 מדדים לצורך בדיקת השיפור בין המודלים:*

*ROC\_AUC – במטרה לראות האם הסיכוי של תיוג חיובי להופיע לפני דוגמא שלילית גדל. כלומר ..*

*F1 – במטרה לראות האם החלוקה האופטימלית של המודל כמסווג בין חיוביים לשליליים השתפרה.*

כיוון ש- דורש *להקשיח את הפרדיקציות בחרנו את סף הציון (להלן ) אשר ממקסם את הציון על הסט הואלידציה (כך ש- ) (זאת כיוון שהבחנו שסף הציון האופטימלי עבור, במקרים רבים, אינו בהכרח 0.5 ולכן בחרנו אותו בשיטה זו ואנו ממליצים למי שמשתמש בשיטות דומות). למען הסר ספק, ביצענו את התהליך הזה עבור כל מודל, לפני ואחרי השימוש בprior.*

*כיוון שיש לנו מס' מחלקות, מיצענו בין המדדים בשיטת 'macro'. השתמשנו בשיטת 'macro' כיוון שיש מס' מחלקות שמרביתן חיוביות ויש מס' מחלקות שמרביתן שליליות, ומצאנו לנכון לבחון אותן ללא תלות במס' החיוביים.*

השתמשנו בשיטת K-Fold Cross Validation כאשר k=5. הציונים המוצגים בתוצאות הינם ממוצע החלוקות השונות.

## חלק I - מודל מרקובי

ראשית, בדקנו את הציון של מודל OneVsRest עבור מודלי בסיס שונים (*logistic, SVM, LightGBM) והשוונו אותם לציון של המודל לאחר הפעלת הפריור. ניתן לראות את התוצאות ב*טבלה 2*.*

|  |  |  |  |
| --- | --- | --- | --- |
| LGB | SVM | Logistic |  |
|  |  |  | OneVsRest |
|  |  |  | OneVsRest and Prior |
|  |  |  | Diffrence |

|  |  |  |  |
| --- | --- | --- | --- |
| LGB | SVM | Logistic |  |
|  |  |  | OneVsRest |
|  |  |  | OneVsRest and Prior |
|  |  |  | Diffrence |

*טבלה 2 – תוצאות עבור חלק 1*

ניתן לראות כי ה-prior משפר את התוצאות עבור המודלים ....

ניתן לראות כי השיפור הניכר ביותר קורה ב-LightGBM. אנחנו חושבים שכך כי .. (מודל יותר אקספרסיבי ולכן ...).

על כן האלגוריתם שהצענו שיפר ביצועים !!!! (לקוראים בעתיד ..)

## חלק II - סוגי עצים

לאחר מכן, רצינו לבדוק כיצד מבנים שונים של מודלים מרקוביים משפיעים על טיב התוצאות.

בנינו מודלים שונים ע"י ואריציות על אלגו' צאו לי .

הסתכלנו על:

1. עץ רנדומלי – בדיקת שפיות שנועדה לבדוק שהשיפורים אכן מגיעים מטיב האלגוריתם.
2. 'Positive' M.I. – נשים לב שנוסחת MI מורכבת מ-4 חלקים מארבעת האפשרויות ללייבל של i,j. רצינו להתחשב אך ורק במקרה שבו שני הלייבלים הם 1 בניסיון לבחון רק על.קורלציה חיובית בין המחלקות תביא לשיפור.

היינו מניחים ש"קורלציה חיובית" תייצג השפעה ואילו קורלציות אחרות אינן רלוונטיות.

Usecase – 2 לייבלים שלא מופיעים כמעט אף פעם כחיובים, בעלי MI גבוה, אמנם לא אמורים להשפיע על הפרדיקציות אחד של השני.

1. 'Semi-Positive' M.I. – 0,1 , 1,0 , 1,1. – אותו הסבר כמו positive, רק שבאמת מעניין אם הראשון בוא 1 והשני 0 ולהיפך.
2. 'Negetive' M.I. - ??

תוצאות ניתן לראות ב*טבלה 3*.

|  |  |  |  |
| --- | --- | --- | --- |
| LGB | SVM | Logistic |  |
|  |  |  | baseline |
|  |  |  | Random |
|  |  |  | M.I. |
|  |  |  | 'Positive' M.I. |
|  |  |  | 'Negetive' M.I. |

|  |  |  |  |
| --- | --- | --- | --- |
| LGB | SVM | Logistic |  |
|  |  |  | baseline |
|  |  |  | Random |
|  |  |  | M.I. |
|  |  |  | 'Positive' M.I. |
|  |  |  | 'Negetive' M.I. |

*טבלה 3 – תוצאות עבור חלק 2*

כפי שניתן לראות ...

## חלק III - הרכבה של מודלים

לבסוף, בדקנו מה קורה כאשר מרכיבים מס' מודלים ..

ההנחה שלנו היא ..

תוצאות ניתן לראות ב*טבלה 4*.

|  |  |  |
| --- | --- | --- |
| LGB with normalization | LGB |  |
|  |  | baseline |
|  |  | M.I. |
|  |  | M.I. -> M.I. |
|  |  | M.I. -> 'Positive' M.I. |
|  |  | M.I. -> 'Positive' M.I. |

*טבלה 4 – תוצאות עבור חלק 3*

ניסינו לנרמל כי משהו על הסתברות (כמו ממקודם \*) נורמליזציה משהו ..

ניתן לראות כי ...

# סיכום ומסקנות

נתקלנו בבעיית Multi Label, עבורה יצרנו מודל אשר מפריד כל תיוג בפי עצמו, אליו הוספנו שימוש ב-prior ע"י מודל מרקובי על המחלקות מבוסס אלגוריתם צ'או ליאו. ראינו שה-prior משפר את תוצאות המודלים, בפרט ובעיקר במודל מבוסס LightGBM. ניסינו ואריציות שונות של אלגוריתם צ'או-ליאו והשוונו אותם עבור מודל עם עץ רנדומלי. ראינו כי שימוש ב-M.I. נותן את התוצאות הטובות ביותר. לבסוף ניסינו להרכיב בין המודלים השונים בשביל להיעזר בשונות בין המודלים לצורך שיפור ה-prior, אמנם ניסיון זה העלה חרס.

1. <https://en.wikipedia.org/wiki/Category:Musicians> [↑](#footnote-ref-1)
2. <https://en.wikipedia.org/wiki/Sean_Tyla> [↑](#footnote-ref-2)