

מעריך שווי, על עץ החלטה ויער אקראי כבר שמעת?

עץ החלטה הוא אלגוריתם של למידה בהשגחה אשר יכול לשמש לחיזוי או לסיווג ויער אקראי הוא הרכב של עצי החלטה. יצאתי לראיין את רועי פולניצר בנושא

עץ החלטה הוא דרך שבה המאפיינים נלקחים בחשבון אחד-אחד. רוצה לומר, בגישת עץ החלטה המאפיינים נלקחים בחשבון אחד אחרי השני לפי סדר חשיבותם, בעוד שבגישת הרגרסיה המאפיינים נלקחים בחשבון בו-זמנית. בנוסף, גישת עץ החלטה איננה מניחה לינאריות, היא הרבה יותר אינטואטיבית והרבה פחות רגישה לתצפיות חריגות מאשר הרגרסיה הלינארית.

מהו מדד אנטרופי?

מדד אנטרופי הוא מדד לאי וודאות. מדד אנטרופי מוגדר באופן הבא, עבור n תוצאות אלטרנטיביות:

$$-\sum_{i=1}^n p_i \ln(p_i)$$

כאשר p_i היא ההסתברות לקבלת התוצאה ה- i .

מהו מדד ג'יני?

מדד ג'יני גם הוא מדד לאי וודאות. מדד ג'יני מוגדר באופן הבא, עבור n תוצאות אלטרנטיביות:

$$1 - \sum_{i=1}^n p_i^2$$

כאשר p_i היא ההסתברות לקבלת התוצאה ה- i .

מהו רווח המידע?

רווח המידע (Information Gain) מוגדר כירידה באנטרופי או במדד הג'יני, הווה אומר, ירידה ברמת אי הוודאות (Information Content).

כיצד אתה בוחר את רמת הסף עבור משתנה נומרי בעץ החלטה?

ערך הסף הינו הערך שממקסם את רווח המידע.

מדען הנתונים רועי פולניצר מכהן כמנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA). רועי בעל תואר M.B.A. במנהל עסקים (עם התמחות בבניית מודלים מתמטיים וסטטיסטיים) ותואר B.A. בכלכלה (עם התמחות בכלים ושיטות לאנליזה ותחקור מידע), שניהם מאוניברסיטת בן-גוריון בנגב ובהצטיינות. רועי למד אקטואריה (בניית מערכות המלצה ואימון רשתות נוירונים לעיבוד תמונה ו-NLP) בכמה מקומות וביניהם בתוכנית ההכשרה היוקרתית בת ה-500 שעות של מכללת John Bryce. בנוסף, רועי חבר באיגוד הבינלאומי למנהלי סיכונים מקצועיים (Professional Risk Managers' International Association), מה שמעיד עליו כבקיא במחקר, פיתוח, כתיבה ומימוש אלגוריתמים של בינה מלאכותית על כמויות גדולות של מידע. בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA - Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

עץ החלטה (Decision Tree) הוא אלגוריתם של למידה בהשגחה אשר יכול לשמש לחיזוי. לעץ החלטה ישנם מספר יתרונות על פני רגרסיה לינארית או רגרסיה לוגיסטית. יתרונו המרכזי הוא שהוא תואם לדרך שבה מרבית בני האדם חושבים על בעיה מסוימת ועל כן הם פשוט וקל להסבר גם למי שאינו מומחה ללמידת מכונה. יתרון אחר של עץ החלטה הוא שאין בו שום דרישה לקשר לינארי בין היעד (Target, המשתנה המוסבר) והמאפיינים (Features, המשתנים המסבירים). יתרון נוסף של עץ החלטה הוא שהעץ בוחר אוטומטית את המאפיינים הטובים ביותר לצורך החיזוי. יתרון אחר של עץ החלטה הוא שהעץ פחות רגישה לתצפיות חריגות מאשר רגרסיה.

רבים חושבים שהמונח עץ החלטה בהקשר של למידת מכונה, פירושו מודל רשתתי (Lattice) כמו המודל הבינומי הבנוי בשיטה של "עץ החלטות", מביא בחשבון מצבי טבע שונים ומאפשר שינוי פרמטרים לאורך התקופה. כיום אני יודע שאין קשר בין העץ הבינומי לבין גישת עץ החלטה. מהם ההבדלים בין גישת עץ החלטה לחיזוי לבין גישת הרגרסיה?

ההלוואה; ציון האשראי של מבקש ההלוואה (FICO).

בעלות על בית, X_1 , 1 = בעלות = 0 = שכירות	הכנסה (באלפי דולר), X_2	יחס החוב להכנסה, X_3	ציון אשראי (FICO), X_4	תוצאת ההלוואה, טובה = 1, חדלת פירעון = 0
1	43.304	18.47	690	0
1	136.000	20.63	670	1
0	38.500	33.73	660	0
1	88.000	5.32	660	1
.....
.....

מהו השלב הראשון אם רוצים ליישם כאן את אלגוריתם עץ ההחלטה?

השלב הראשון הוא לחשב את רווח המידע הצפוי (הירידה באנטרופי הצפוי) מכל אחד מהמאפיינים. מתוך מבקשי ההלוואה, 59.14% בבעלותם בתים בעוד ש- 40.86% שוכרים. ההלוואות היו טובות עבור 84.44% ממבקשי ההלוואות שהיו בבעלותם בתים ו- 80.33% מאלו ששכרו. לפיכך, אם ידוע שבבעלותו של מבקש ההלוואה בית הרי שהאנטרופי הצפוי הופך ל-:

$$0.5914 \times [-0.8444 \times \ln(0.8444) - 0.1556 \times \ln(0.1556)] + 0.4086 \times [-0.8033 \times \ln(0.8033) - 0.1967 \times \ln(0.1967)] = 0.4582$$

הירידה הצפויה באנטרופי היא מועטה

$$0.4597 - 0.4582 = 0.0014$$

החישוב של האנטרופי הצפוי מתוך ההכנסה דורש הגדרת רמת סף של הכנסה. לפיכך נגדיר:

p_1 : ההסתברות שההכנסה של הלווה גבוהה יותר מרמת הסף

p_2 : ההסתברות שאם ההכנסה של הלווה גבוהה יותר מרמת הסף, הלווה לא יגיע לחדלות פירעון

p_3 : ההסתברות שאם ההכנסה של הלווה נמוכה יותר מרמת הסף, הלווה לא יגיע לחדלות פירעון

האנטרופי הצפוי הוא

$$P_1 \times [-P_2 \times \ln(P_2) - (1 - P_2) \times \ln(1 - P_2)] + (1 - P_1) \times [-P_3 \times \ln(P_3) - (1 - P_3) \times \ln(1 - P_3)]$$

ביצעתי חיפוש איטרטיבי על מנת לקבוע את הערך של רמת הסף אשר ממקסמת את האנטרופי הצפוי עבור סט האימון. מתברר שרמת הסף היא \$85,193. עבור ערך זה של רמת הסף, $p_1=29.93\%$, $p_2=87.82\%$ ו- $p_3=80.60\%$.

להלן התוצאות עבור כל החישובים של רווח המידע:

מאפיין	ערך הסף	אנטרופי צפוי	רווח המידע
Home Ownership	N.A.	0.4582	0.0014
Income	\$85,193	0.4556	0.0040
Debt to Income (dti)	19.87	0.4576	0.0021
FICO	716	0.4536	0.0061

הבנתי שלא רק שיש למידה בהשגחה, למידה בהשגחה למחצה ולמידה ללא השגחה, הרי שגם יש למידת הרכב. מהי למידת הרכב?

למידת הרכב (Ensemble Learning) הינה למידה באמצעות שילוב תוצאות מכמה אלגוריתמים.

מהי שיטת הרכב?

שיטת הרכב (Ensemble Method) הינה דרך מסוימת לשלב מספר אלגוריתמים לצורך חיזוי בודד.

מהו יער אקראי?

יער אקראי (Random Forest) הוא הרכב של מספר עצי החלטה. עצי ההחלטה השונים נוצרים באמצעות שימוש בתתי קבוצות (Subset) של מאפיינים או בתתי קבוצות של תצפיות או על ידי שינוי ערכי הסף.

מהי טכניקת ה-Bagging?

Bagging היא טכניקה של אימון אותו האלגוריתם על תתי קבוצות של נתונים מקריות שונות.

מהי טכניקת ה-Boosting?

Boosting היא פרוצדורת אימון איטרטיבית שבמסגרתה אלגוריתם אחד מנסה לתקן את השגיאות של האלגוריתם שקדם לו.

האם תוכל להסביר את ההבדל בין Bagging ו-Boosting?

בוודאי. בעוד ש-Bagging הינה טכניקה של דגימה מקרית מתוך תצפיות או מאפיינים כך שאותו אלגוריתם משמש על סטים שונים של נתוני אימון, הרי ש-Boosting היא טכניקה של יצירת מודלים ברצף כאשר כל אחד מהמודלים מנסה לתקן את הטעויות של המודל שלפניו.

קראתי באחד המאמרים שלך שכתבת שאחד היתרונות של אלגוריתם עץ ההחלטה הוא שהוא שקוף, למה התכוונת?

כוונתי הייתה שהאלגוריתם עץ ההחלטה הוא שקוף בכך שהוא מאפשר בקלות יחסית לראות מדוע החלטה מסוימת התקבלה.

האם תוכל לתת איזשהי דוגמה ליישום של עץ החלטה עבור החלטות אשראיי?

ניקח לדוגמה סט נתונים של חברת Prosper Marketplace לגבי החלטות האשראי שלה. Prosper Marketplace היא מלווה מסוג peer-to-peer (הלוואות חברתיות) המאפשרת למשקיעים להלוות כסף ללווים ללא תיווך. Prosper Marketplace משתמשת בלמידת מכונה ומפרסמת נתונים על הלוואותיה. ניסיתי ליישם את גישת עץ ההחלטה באמצעות מדד האנטרופי. סט האימון מורכב מ- 8,695 תצפיות וסט הבדיקה מורכב מ- 5,916 תצפיות. כמובן שבסט האימון 7,196 היו הלוואות טובות ו- 1,499 הלוואות שהגיעו לחדלות פירעון. ללא כל מידע נוסף, ההסתברות האמפירית להלוואה טובה היא 7,196/8,695 או 82.76%. לפיכך, האנטרופי ההתחלתי הוא:

$$0.8276 \times \ln(0.8276) - 0.1724 \times \ln(0.1724) = 0.4597$$

סט האימון וסט הבדיקה מורכבים מארבעת המאפיינים הבאים: (1) משתנה קטגוריאלי המציון האם יש בבעלותו של מבקש ההלוואה בית או שמא הוא שוכר; (2) הכנסתו של מבקש ההלוואה; (3) יחס החוב להכנסה של מבקש

או אז ההלוואה מסווגת כטובה ואם $Z \leq Q$ או אז ההלוואה מסווגת כלא טובה. ניתן לסכם את התוצאות המתקבלות מיישום ערך מסוים של Z על סט הבדיקה, באמצעות מה שמכונה מטריצת טעות (Confusion Matrix). מטריצת הטעות מציגה את הקשר שבין הסיווגים לבין התוצאות בפועל.

מטריצת הטעות עבור סט הבדיקה כאשר $Z = 0.75$		
סיווג כחדלת פירעון	סיווג כלא חדלת פירעון	
13.66%	68.46%	תוצאה חיובית (אין חדלות פירעון)
4.19%	13.69%	תוצאה שלילית (חדלות פירעון)

מטריצת הטעות עבור סט הבדיקה כאשר $Z = 0.8$		
סיווג כחדלת פירעון	סיווג כלא חדלת פירעון	
29.39%	52.72%	תוצאה חיובית (אין חדלות פירעון)
8.71%	9.18%	תוצאה שלילית (חדלות פירעון)

מטריצת הטעות עבור סט הבדיקה כאשר $Z = 0.85$		
סיווג כחדלת פירעון	סיווג כלא חדלת פירעון	
47.01%	35.11%	תוצאה חיובית (אין חדלות פירעון)
12.53%	5.36%	תוצאה שלילית (חדלות פירעון)

לקחתי ערכי Z של 0.75, 0.8 ו-0.85. ניתן לראות בתרשים העץ לעיל את הקשרים שבין ערך ה- Z לבין ההחלטות המתקבלות. $Z=0.75$ מנבה שכל ההלוואות הן טובות מלבד אלו שעבורן $FICO \leq 716$, $Income \leq \$85,240$, $dti > 16.54$ ושהלווה שוכר דירה. לחילופין, $Z=0.80$ מנבה שכל ההלוואות הן טובות מלבד אלו שעבורן $FICO \leq 716$, $Income \leq \$85,240$ ו- $dti > 16.54$. לחילופין, $Z=0.85$ מנבה שכל ההלוואות הן טובות מלבד אלו שעבורן $FICO \leq 716$ ו- $Income \leq \$85,240$.

האם אתה יכול בבקשה לסדר את ההגדרות הללו במטריצת הטעות על מנת שבנין יותר טוב כיצד הן משתלבות?

סיווג כתוצאה חיובית	סיווג כתוצאה שלילית	
TP	FN	תוצאה חיובית
FP	TN	תוצאה שלילית

להלן היחסים הנגזרים ממטריצת הטעות:

$$\text{The False Positive Rate} = \frac{FP}{TN + FP}$$

$$\text{The False Negative Rate} = \frac{FN}{TP + FN}$$

$$\text{True Positive Rate} = \frac{TP}{TP + FN}$$

$$\text{The True Negative rate} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

ניתן לראות שלציון ה- $FICO$ יש את רווח המידע הגבוה ביותר. לפיכך, מיקמתי את ציון ה- $FICO$ בצומת השורש (Root Node) של העץ. הענפים ההתחלתיים של העץ תואמים ל- $FICO > 716$ ו- $FICO \leq 716$.

מה אתה עושה ביתר העץ?

עבור הרמה הבאה של העץ אני חוזר על התהליך. עבור $FICO > 716$ ו- $FICO \leq 716$, חישבתי את רווח המידע עבור כל אחד משלושת המאפיינים הנותרים. להלן תוצאות החישובים הללו:

רווח המידע של המאפיינים לקביעת הרמה השנייה של העץ				
הענף הראשון	מאפיין	ערך הסף	אנטרופי צפוי	רווח המידע
$FICO > 716$	Home Ownership	N.A.	0.3050	0.0002
$FICO > 716$	Income	\$48,711	0.3001	0.0050
$FICO > 716$	Debt to Income (dti)	21.13	0.3035	0.0016
$FICO \leq 716$	Home Ownership	N.A.	0.4870	0.0012
$FICO \leq 716$	Income	\$85,244	0.4844	0.0038
$FICO \leq 716$	Debt to Income (dti)	16.80	0.4536	0.0061

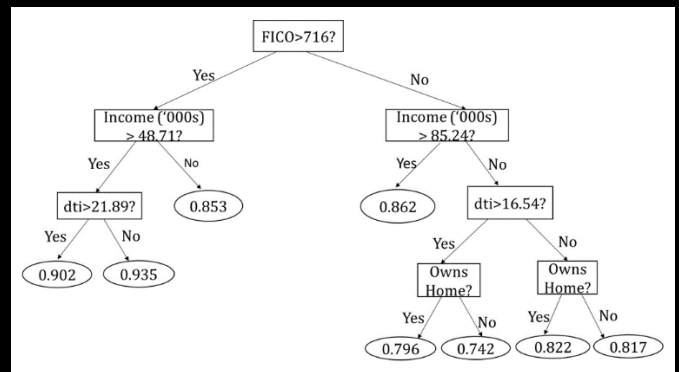
מתברר שההכנסה היא המאפיין הבא שעלינו להתחשב בו, הן כאשר $FICO > 716$ והן כאשר $FICO \leq 716$.

האם ככה זה תמיד באלגוריתם עצי החלטה שהמאפיין הבא הנבחר יהיה זה עבור שני הענפים שיוצאים מאותו הצומת?

ממש לא. ככה יצא במקרה דנן שלפנינו אבל אי אפשר לומר שלרוב זה מה שיוצא. בנוסף, כאשר המאפיין הבא הנבחר הינו זהה עבור שני הענפים, אז על פי רוב לא יהיה לו את אותו ערך הסף עבור שני הענפים. בדוגמה שלי, כאשר $FICO > 716$ הענפים העוקבים התואמים למקרים שלנו הינם: $Income > \$48,711$ ו- $Income \leq \$48,711$. כאשר $FICO \leq 716$, הענפים העוקבים התואמים למקרים הם $Income > \$85,244$ ו- $Income \leq \$85,244$.

כיצד נראה העץ המלא?

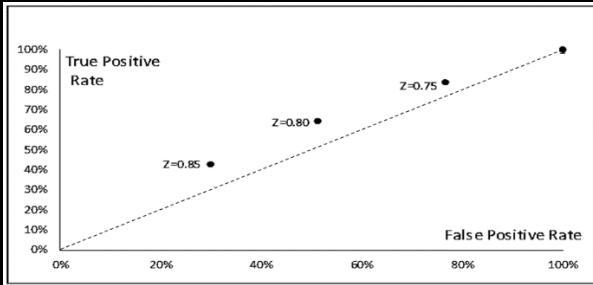
להלן העץ המלא:



בעת בניית העץ לעיל "גזזתי" אותו וביטלתי נקודות החלטה המתאימות לפחות מ-1,000 תצפיות וכאלו שבהן לאחת התוצאות היו פחות מ-200 תצפיות. הערה אינפורמטיבית: ניתן להשתמש במבחנים סטטיסטיים כמו מבחן חי-בריבוע (chi-squared test) על מנת לקבוע האם רווח המידע מובהק סטטיסטית. הנקודות הסופיות של העץ המוצגות כאליפסות, מכונות "עלים" (Leaves). המספרים המופיעים בעלים מייצגים את ההסתברות להלוואה טובה המתקבלת מסט האימון. כעת על מדען הנתונים להחליט על קריטריון לסיווג האם הלוואה מסוימת הינה טובה או לא טובה. ההחלטה על קריטריון כרוכה בהגדרת רמת סף, Z , עבור הערך של Q כך שאם $Q > Z$

לפני שהצגת את הדוגמא הזכרת את עקומת ה-ROC. האם תרצה להרחיב עליה?

אני אציג תרשים שמתאר את שיעור החיוביים האמיתיים מול שיעור השליליים הכוזבים. תרשים זה מכונה עקומת ה-ROC.



Z = 0.85	Z = 0.80	Z = 0.75	
29.96%	51.32%	76.56%	שיעור החיוביים הכוזבים (FPR)
57.25%	35.79%	16.63%	שיעור השליליים הכוזבים (FNR)
42.75%	64.20%	83.37%	שיעור החיוביים האמיתיים (TPR)
70.04%	48.68%	23.44%	שיעור השליליים האמיתיים (TNR)
47.63%	61.43%	72.65%	נכונות (Accuracy)
86.76%	85.17%	83.33%	דיוק (Precision)

מה פירוש שיעור החיוביים האמיתיים (True Positive Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כחיובית מותנה בכך שידוע שהתוצאה היא חיובית. בסטטיסטיקה מדד זה מכונה רמת ביטחון או רמת סמך. לדוגמא, ההסתברות לסווג הלואה מסוימת כלא חדלת פירעון אם ידוע שלא ארעה חדלות פירעון.

השטח שמתחת לעקומה (AUC) הינו דרך פופולארית לסכם את יכולת הסיווג של המודל. אם ה-AUC הוא 1.0, אז המודל הוא מושלם היות ושיעור החיוביים האמיתיים הוא 100% ושיעור השליליים הכוזבים הוא 0%. הקו המנוקד שבתרשים לעיל מתאים ל-AUC של 0.5, מה שמתאים למודל ללא יכולת סיווג. למשל, למודל שמבצע סיווג מקרי יש AUC של 0.5.

מה פירוש שיעור החיוביים הכוזבים (False Positive Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כחיובית מותנה בכך שידוע שהתוצאה היא שלילית. בסטטיסטיקה מדד זה מכונה רמת המובהקות או אלפא. לדוגמא, ההסתברות לסווג הלואה מסוימת כלא חדלת פירעון אם ידוע שארעה חדלות פירעון.

אז מודלים עם AUC שקטן מ-0.5 הם למעשה גרועים יותר ממודלים שמבצעים סיווג מקרי. מה לגבי המודל שלך?

מהו מדד שיעור השליליים האמיתיים (True Negative Rate)?

זוהי ההסתברות המותנה לסווג תצפית מסוימת כשלילית מותנה בכך שידוע שהתוצאה היא שלילית. בסטטיסטיקה מדד זה מכונה עוצמת המבחן. לדוגמא, ההסתברות לסווג הלואה מסוימת כחדלת פירעון אם ידוע שארעה חדלות פירעון.

עבור הנתונים שבחנתי למודל עץ ההחלטה יש יכולת סיווג נמוכה. בהינתן שחברת Prosper Marketplace כבר משתמשת בלמידת מכונה לצורך קבלת החלטות ההלוואה שלה ושאי משתמש רק בארבעה מאפיינים, אז אין זה מפתיע שה-AUC של המודל שלי הוא רק מעט מעל ל-0.5. הנקודה החשובה היא שאין לצפות שמודל שכזה יבצע סיווג מושלם.

מהו מדד שיעור השליליים הכוזבים (False Negative Rate)?

לאמור- זוהי ההסתברות המותנה לסווג תצפית מסוימת כשלילית מותנה בכך שידוע שהתוצאה היא חיובית. בסטטיסטיקה מדד זה מכונה טעות מסוג 2 או ביתא. לדוגמא, ההסתברות לסווג הלואה מסוימת כחדלת פירעון אם ידוע שלא ארעה חדלות פירעון.

אז מהו המבחן העיקרי עבור מודל סיווג?

המבחן העיקרי עבור מודל הסיווג הוא האם המודל מסוגל לקבל החלטות שטובות לפחות כמו ההחלטות שהיה מקבל בנאדם. רוצה לומר שבעת ההחלטה על הערך הראוי של Z (קרי, המיקום על עקומת ה-ROC) על הלווה לשקול הן את הרווח הממוצע מהלוואות שלא מגיעות לחדלות פירעון והן את ההפסד הממוצע מההלוואות שמגיעות לחדלות פירעון.

מהו מדד הנכונות (Accuracy)?

נכונות הוא אחוז התצפיות שסווגו נכונה.

רגע, אז מה קורה אם הרווח מהלוואה שלא מגיעה לחדלות פירעון הוא X, בזמן שההפסד מהלוואה שמגיעה לחדלות פירעון הוא 4X?

מהו מדד הדיוק (Precision)?

מדד הדיוק הוא אחוז הסיווגים החיוביים אשר היו נכונים. לאמור- זוהי ההסתברות המותנה שתוצאה מסוימת היא חיובית מותנה בכך שידוע שהתצפית סווגה כחיובית. בסטטיסטיקה מדד זה מכונה ערך ניבוי חיובי. לדוגמא, ההסתברות שהלוואה מסוימת לא תגיע לחדלות פירעון אם ידוע שהיא סווגה כלא חדלת פירעון.

אז ישנם למעשה מספר יחסי תחלופה.

אז הרווח של המלווה הוא גבוה ביותר כאשר הוא ממקסם את הפונקציה הבאה:

$$X \times TP - 4X \times FP$$

עבור האלטרנטיבות שבדקנו לעיל (Z של 0.75, 0.8, ו-0.85) הפונקציה הבאה שווה שווה 13.70X, 16.00X, ו-13.67X בהתאמה בהתאמה. זה מצביע על כך שמשלושת ערכי ה-Z האלטרנטיביים, Z = 0.8 הוא הרווחי ביותר.

אני מאחל לך הצלחה רבה בהתמודדות על תפקיד האקטואר הראשי באוצר.

בדיוק כך. אני יכול להגדיל את שיעור השליליים האמיתיים (קרי, לזהות אחוז גבוה יותר של הלוואות שיגיעו לחדלות פירעון) רק אם אני אזהה אחוז נמוך יותר של הלוואות שהוכחו כטובות. כמובן שמדד הנכונות יורד ככל ששיעור השליליים האמיתיים עולה.

תודה רבה.