פרק 8: עצי החלטה

עצי החלטה הינם כלי נפוץ ופשוט יחסית לסיווג ורגרסיה. בבסיסו עץ החלטה הינו מימוש מסוים של פונקציות לוגיות (כניסה ויציאה דיסקרטית), אולם ניתן ליישמו בו גם עבור משתנים רציפים על ידי דיסקרטיזציה. קיימים מספר אלגוריתמים סטנדרטיים (ותוכנות מתאימות) לבניית עצי החלטה מדוגמאות: הנפוצים ביותר נקראים CART (Classification and Regression Trees) ו-C5.0 (לשעבר C4.5, ID3). יישומים רבים קיימים בתחומים כגון אבחון רפואי, חיזוי פיננסי, כריית מידע עסקי, ועוד.

נעמוד פה בקצרה על העקרונות הבסיסיים של בניית עצי החלטה.

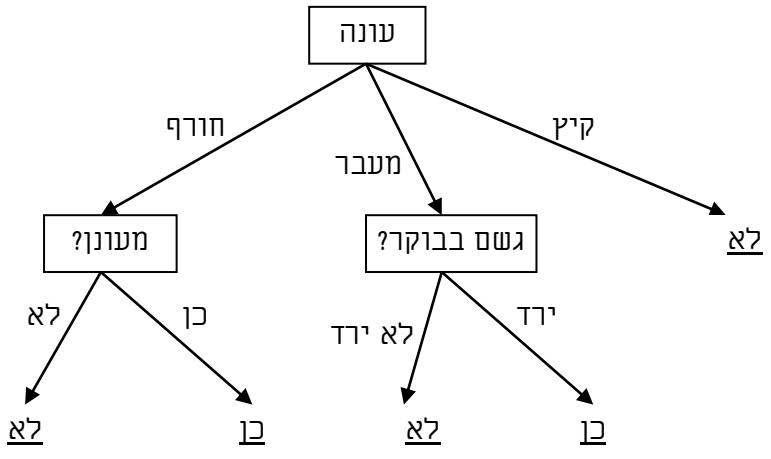
8.1 מבנה עצי החלטה

נבהיר מהו עץ החלטה באמצעות דוגמא פשוטה.

נניח כי עלינו להעריך באם יֵרד בצהרים גשם, על סמך המאפיינים (attributes) הבאים, שנצפו בבוקר אותו יום:

- עונה: קיץ/חורף/מעבר
- יָרד גשם בבוקר?: כן/לא
- מעונן?: כן/לא

מוצע עץ ההחלטה הבא:



ציור 1: עץ החלטה עבור ההשערה "יֵרד גשם בצהרים"

עצי החלטה מסווגים את המקרים הנתונים החל משורש העץ (המאפיין "עונה") וכלה בעלים (ההחלטות "*כן*" ו"*לא*"). כל ענף היוצא מצומת כלשהו מתייחס לערך שונה של המאפיין המוצב בצומת זה, ומוביל לצומת נוסף או ל"עלה" החלטה. לשאלה הנמצאת בכל אחת מהצמתים

נתייחס כ"מבחן" המוצב בצומת זה. מבחן זהה לא יופיע יותר מפעם אחת בכל "ערוץ" מהשורש לעלה; אולם הוא עשוי להופיע מספר פעמים בערוצים שונים.

עץ ההחלטה הנ"ל שקול לפונקציה הלוגית הבאה:

$$((S = \text{מעבר}) \wedge (R = \text{כן})) \vee ((S = \text{חורף}) \wedge (C = \text{כן}))$$

כאשר S = עונה, R = גשם בבוקר, C = מעונן.

כללית, כל עץ החלטה בעל תוצאה בינארית (True/False) מייצג פונקציה לוגית בצורת Disjunctive Form Normal" (ביטויי "AND" מקושרים ע"י פעולות "OR"). קל לראות כי ניתן לייצג כל פונקציה לוגית בצורה זו. לפיכך, הייצוג ע"י עצי החלטה הינו כללי.

דוגמא נוספת לעץ החלטה המשמש לזיהוי פירות מוצגת בציור הבא. במקרה זה התוצאה אינה בינארית.

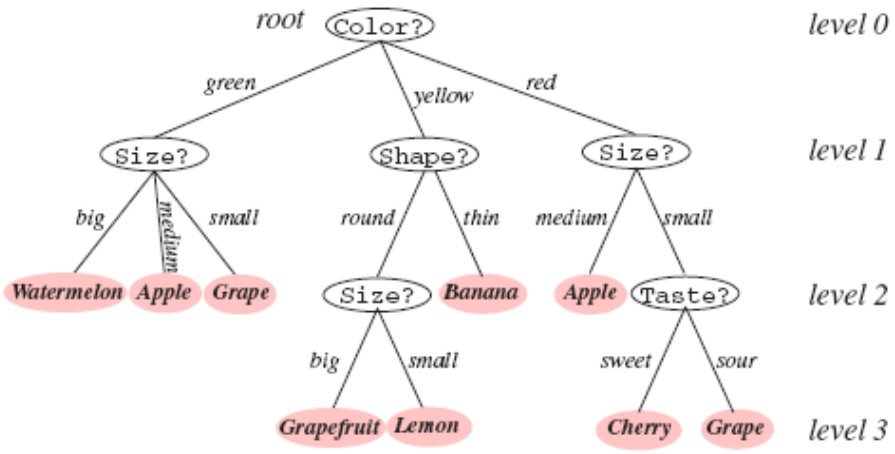
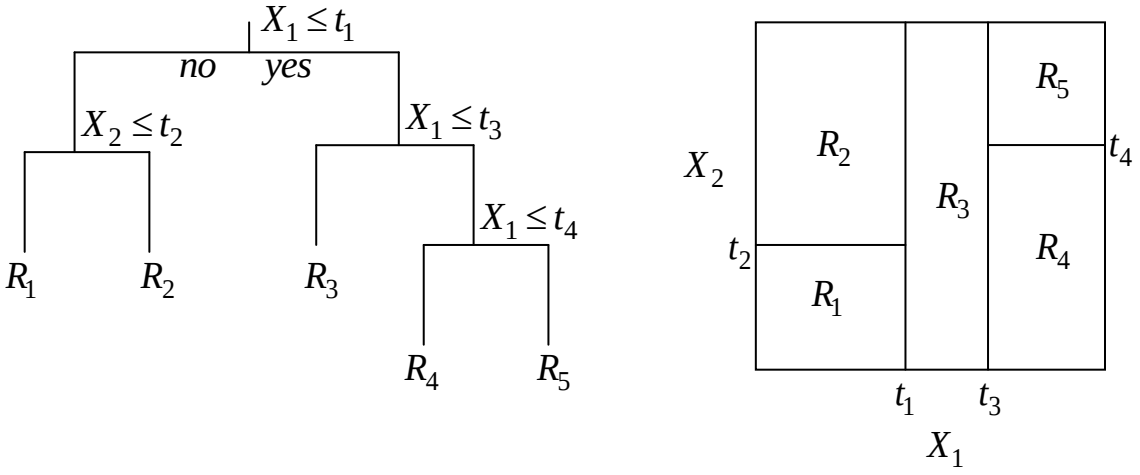


FIGURE 8.1. Classification in a basic decision tree proceeds from top to bottom. The questions asked at each node concern a particular property of the pattern, and the downward links correspond to the possible values. Successive nodes are visited until a terminal or leaf node is reached, where the category label is read. Note that the same question, `Size?`, appears in different places in the tree and that different questions can have different numbers of branches. Moreover, different leaf nodes, shown in pink, can be labeled by the same category (e.g., **Apple**). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

הדוגמא הבאה (לפי HTF:2001) מראה שימוש בעץ החלטה לחלוקת מרחב כניסה רציף (משתני כניסה ממשיים X_1, X_2) למספר אזורים מלבניים (R_ℓ). חלוקה זו מאפשרת לממש את הפונקציה הבאה:

$$f(x) = \sum_{\ell=1}^{M} c_\ell I\{x \in R_\ell\}$$

דהיינו פונקציה קבועה-למקוטעין המקבלת ערך c_i באזור R_i של מרחב הכניסה.



ציור 3 : חלוקה של מרחב מאפיינים דו-מימדי ורציף על ידי פיצול בינארי חוזר (בהתאם ל- CART)

8.2 בנית עצי החלטה

נראה עתה כיצד ניתן לבנות עץ החלטה לבעיית סיווג מתוך אוסף דוגמאות נתון. כזכור סדרת הדוגמאות היא מהצורה $\{x_k, d_k\}_{k=1}^{n}$, x הוא וקטור המאפיינים, והתווית d מציינת את הערך הרצוי במוצא. וקטור הקלט x עשוי לכלול מספר רב של מאפיינים, שלחלקם רלוונטיות נמוכה לתוצאה המבוקשת. כמו כן ייתכן אחוז מסוים של דוגמאות שגויות.

המטרה הכללית היא לבנות על סמך הדוגמאות עץ החלטה בעל התכונות הבאות :

1. סיווג נכון של מרבית הדוגמאות.
2. עץ קצר (פשוט) ככל הניתן.

תכונה 2 חשובה משתי סיבות :

1. פשטות המימוש.
2. יכולת הכללה : מניעת התאמת-יתר לאוסף הדוגמאות הנתון.

באופן כללי יותר, העדפת עצי-החלטה קצרים קשורה בנקודות הבאות :

1. "Occam's Razor" : מבין מספר הסברים אפשריים לתופעה נתונה, יש להעדיף את הפשוט ביותר.

2. העדפה מעין זו (וליתר דיוק, אופן ביטויה באלגוריתם המסוים שממומש) מאפשרת בחירה של עץ החלטה מסוים מתוך מספר רב של עצים אפשריים המתאימים לדוגמאות. העדפה זאת מכונה גם ה"הטיה האינדוקטיבית" (Inductive Bias) של האלגוריתם.

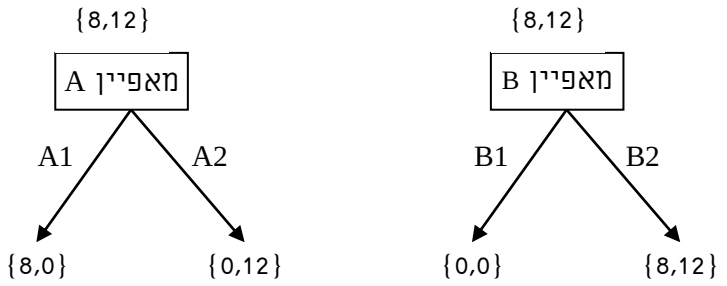
בנית העץ מתמקדת בשאלות הבאות: (1) בחירת המבחנים (המאפיינים) ומיקומם בעץ (2) בחירת התווית של כל עלה בעץ. האלגוריתמים המקובלים מאופיינים על ידי התכונות הבאות:

1. גישת top-down: בנית העץ מראשו (השורש) למטה.
2. אופטימיזציה חד-שלבית (greedy/myopic search): המבחן הבא נבחר בכל שלב תוך הסתכלות של צעד אחד בלבד קדימה (במורד העץ).

א. בחירת המבחן המיטבי: דוגמאות

נתחיל בבחירת המבחן הראשון, הנמצא בשורש העץ. ברצוננו לבחור במבחן אשר נותן, כשלעצמו, סיווג "טוב ככל האפשר" של הדוגמאות. לצורך זה עלינו להגדיר מדד כמותי של סיווג מיטבי. בשלב זה נזהה, לצורך פשטות, את קבוצת המבחנים האפשריים עם קבוצת המאפיינים.

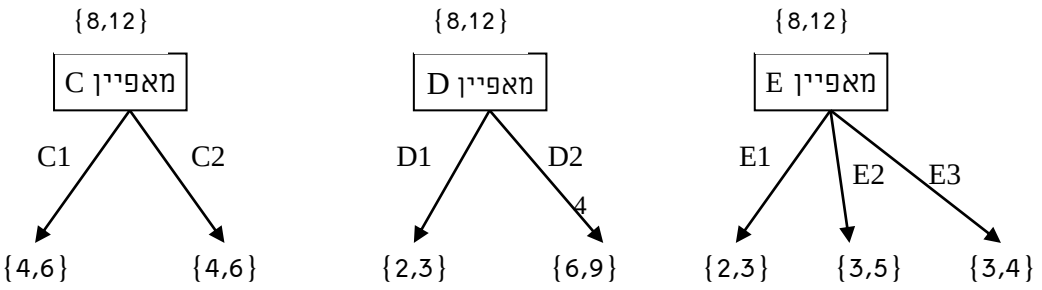
נתחיל במספר דוגמאות להמחשה. עבור בעיית סיווג לשתי מחלקות $\Omega=\{c_1,c_2\}$, נניח כי נתונות 20 דוגמאות, מתוכן $n_1=12$ מסווגות c_1 ($d_k=c_1$) ו-$n_2=8$ מסווגות c_2. נסמן $\{n_1,n_2\}=\{12,8\}$. חלוקת הדוגמאות לפי שני מאפיינים אפשריים היא כלהלן:



ציור 4: חלוקה לפי מאפיינים B,A.

ברור כי מאפיין A הוא האידיאלי: הוא מבצע סיווג מושלם של הדוגמאות. מאפיין B הוא הגרוע ביותר: הוא אינו מוסיף כל מידע חדש.

נתבונן בחלוקה המתקבלת עבור שלושה מאפיינים נוספים:



ציור 5 : חלוקה לפי מאפיינים E, D, C .

מאפיין C ו- D שומרים על פרופורציה זהה של הדוגמאות המסווגות (c_1 לעומת c_2), אך גודל הקבוצות שונה. את מי נעדיף?

מאפיין E נותן חלוקה ל- 3 קבוצות. מכיוון שמדובר בתת-חלוקה (עידון) של החלוקה ע"י מאפיין D, הרי נראה ש- E עדיף על D (יש תוספת אינפורמציה). אולם איך נשווה בין E ל- C ?

ההשוואה והבחירה בין המאפיינים תבוצע כלהלן :

1. נגדיר מדד של "תוספת מידע", אשר ימדוד את המידע המועיל שנוסף לאחר הסיווג ע"י מאפיין נתון.
2. נבחר את המאפיין הנותן תוספת מידע מירבית.

ב. מדדים לחוסר אחידות (impurity)

נתבונן בקבוצת הדוגמאות המסווגות $S = \{x_k, d_k\}_{k=1}^{N}$ שהתקבלה בצומת כלשהו של העץ. נניח כי $d_k \in \{c_1, \ldots, c_C\}$. המטרה הסופית הינה כזכור לקבל בכל צומת דוגמאות "אחידות" (הומוגניות) ככל האפשר מבחינת סיווגן, כלומר בעלות סיווג d_k זהה. נגדיר עתה מספר מדדים אפשריים לחוסר-אחידות של סדרת דוגמאות זו. לשם כך נגדיר ראשית את השכיחות היחסית (או "הפילוג האמפירי") של כל אחד מהסיווגים האפשריים (c_j) בקבוצת הדוגמאות :

$$\hat{p}_j = \frac{1}{N}\sum_{k=1}^{N} I\{d_k = c_j\}$$

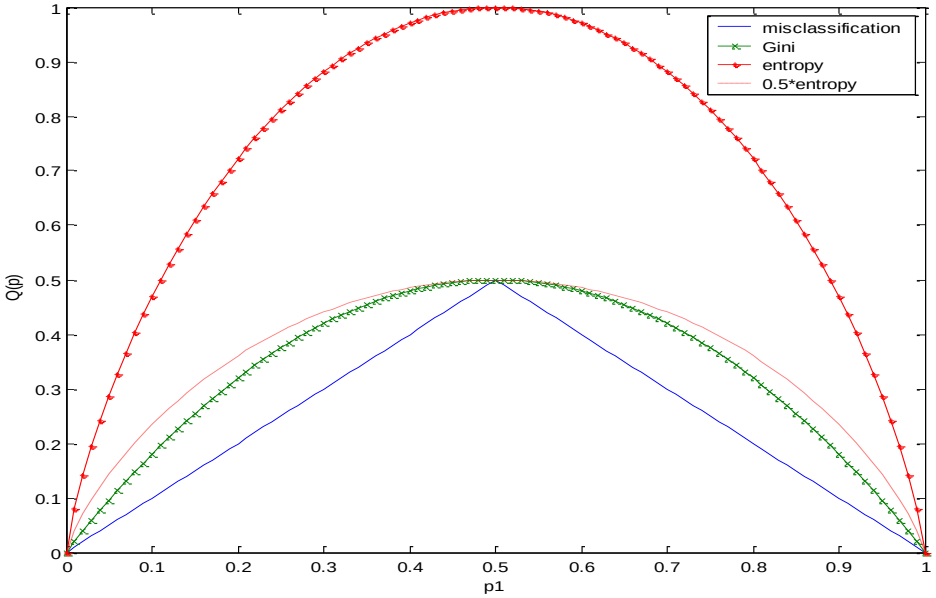
קיימים מספר מדדים מקובלים לחוסר-האחידות של תוויות קבוצת הדוגמאות, שכולם פונקציה של וקטור הפילוג האמפירי $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_C)$:

1. שגיאת הסיווג : $Q(\hat{p}) = 1 - \max_{j \in \{1,\ldots,N\}} \hat{p}_j$
2. אינדקס Gini : $Q(\hat{p}) = \sum_j \hat{p}_j (1 - \hat{p}_j)$

3. אנטרופיה:
$$Q(\hat{p}) = H(\hat{p}) = \sum_j \hat{p}_j \log_2 \frac{1}{\hat{p}_j} = -\sum_j \hat{p}_j \log_2(\hat{p}_j)$$

כל שלושת המדדים מקיימים את התכונות הבאות (ראה ציור עבור המקרה $C = 2$):

1. $Q(\hat{p}) = 0$ עבור פילוג חד-ערכי ($p_j = 1$ עבור j כלשהו).
2. $Q(\hat{p})$ מקבל את ערכו המכסימלי עבור פילוג אחיד ($p_j \equiv 1/C$).



ציור 6: מדדי חוסר-אחידות עבור סיווג בינארי ($C = 2$).

הסימון $Q(S)$ ישמש אותנו בהמשך כסימון חליפי ל- $Q(\hat{p})$.

ג. תוספת המידע של מאפיין

נתבונן בצומת כלשהו בעץ, אליו הגענו עם קבוצה S של N דוגמאות מסווגות. מטרתנו לבחור את המאפיין (או, באופן כללי יותר, המבחן) שיוצב בצומת זה, כהמשך תהליך גידול העץ.

נחשב ראשית את $Q(S)$: מדד חוסר-האחידות עבור קבוצת הדוגמאות S.

נתבונן עתה במבחן A כלשהו, אשר מפצל את קבוצת הדוגמאות S ל- M תת-קבוצות: $\{S_m : m = 1, 2, \ldots, M\}$. עבור כל תת-קבוצה S_m נחשב את מדד חוסר האחידות $Q(S_m)$.

מדד חוסר-האחידות המשוקלל עבור האוסף $\{S_m\}$ יוגדר עתה על ידי:

$$Q(S \mid A) = \sum_{m=1}^{M} \frac{|S_m|}{N} Q(S_m)$$

מדד הטיב של המבחן A ביחס לקבוצת הדוגמאות S יוגדר עתה על ידי

$$\Delta Q(S \mid A) = Q(S) - Q(S \mid A)$$

ניתן לראות כי זהו הגידול באחידות (או הקטנה בחוסר-האחידות) של האוסף $\{S_m\}$ לעומת קבוצת הדוגמאות המקורית S. כאשר $Q(\cdot)$ הינו האנטרופיה, $\Delta Q(S \mid A)$ נקרא גם *תוספת המידע* (information gain) של המבחן A.

המבחן A שנבחר הוא (כעיקרון) זה שעבורו השיפור $\Delta Q(S \mid A)$ הינו מקסימלי.

תרגיל: הראו כי $0 \le Q(S \mid A) \le Q(S)$ (עבור מדד האנטרופיה).

ד. מבנה האלגוריתם הבסיסי

האלגוריתם הבסיסי לבניית העץ הינו כלהלן:

1. העץ נבנה החל מהשורש (המאפיין העליון) כלפי מטה.
2. בכל שלב נבחר את המבחן הנותן "הגדלת אחידות" (או "תוספת מידע") מכסימלית.
3. הבחירה ממשיכה בכל ערוץ במורד העץ עד למילוי תנאי סיום: למשל סיווג מושלם של כל הדוגמאות או מיצוי כל המאפיינים.
4. בכל 'עלה' של העץ שהתקבל, תווית הסיווג נקבעת לפי הרוב.

* ה. גורם הפיצול

בקריטריון שבו השתמשנו קיימת העדפה מובנית למבחנים (מאפיינים) בעלי מספר ערכים רב. בדוגמת הגשם בה פתחנו, עם יעמוד לרשותנו גם התאריך כמאפיין אפשרי, הרי נוכל לסווג את כל הדוגמאות באופן מושלם רק על סמך "מאפיין" זה, והוא בוודאי ייבחר בצומת הראשונה לפי קריטריון "תוספת המידע". אולם לקריטריון זה ערך מועט לצורך חיזוי.

דרך אחת (המקובלת למשל ב-CART) להתמודד עם בעיה זו היא לקבוע מראש כי מספר הפיצולים בכל צומת (גורם הפיצול) יהיה 2. אין בכך כדי לפגוע בכוח הייצוג של העץ כיוון שמבחן בעל $M > 2$ ערכם אפשריים ניתן לממש על ידי $(M-1)$ צמתים בינאריים. מובן שהעץ המתקבל עשוי להיות גדול יותר.

כאשר מספר הפיצולים אינו מוגבל כך, ניתן למנוע את ההטיה לטובת מבחנים בעלי מקדם פיצול גבוה על ידי נרמול "תוספת המידע" של מאפיין A באופן הבא:

$$\Delta\tilde{Q}(S \mid A) = \frac{\Delta Q(S \mid A)}{Split(S,A)}$$

כאשר $Split(S,A)$ הינו מקדם פיצול מתאים. הגדרה מקובלת לגורם פיצול זה הינה

$$Split(S,A) = \log n(A)$$

כאשר $n(A)$ הינו מספר הערכים השונים של המבחן A המתקבלים על פני איברי הקבוצה S.

הגדרה עדינה יותר, המתחשבת גם בגודל היחסי של הקבוצות $\{S_m\}$ המתקבלות, היא כלהלן:

$$Split(S,A) = -\sum_{m=1}^{M} \frac{|S_m|}{|S|} \log \frac{|S_m|}{|S|}$$

הגדרה זו מתלכדת עם הקודמת כאשר כל תת-הקבוצות $\{S_m\}$ שוות בגודלן.

* ו. מאפיינים חסרים

בפועל, חלק מהדוגמאות עשויות להיות בעלות מאפיינים חסרים. למשל: לחולה מסוים לא נמדד לחץ הדם. הדברים אמורים הן לגבי סדרת הלימוד והן לגבי תבניות הקלט שנדרש לסווגן.

דרך פשוטה לטפל בדוגמאות אלו הינה להעניק למאפיין החסר את הערך של רוב הדוגמאות בצומת שבה מאפיין זה נבחן. סכמות אחרות מגרילות את הערך החסר לפי פילוג הדוגמאות באותה צומת. נעיר שיש אינפורמציה בכך שמאפיין מסוים לא נמדד (אי מדידת לחץ דם יכולה לתת אינדיקציה לכך הרופא לא ראה לנכון למדוד את לחץ הדם ולכן יש סבירות גבוהה יותר ללחץ דם תקין).

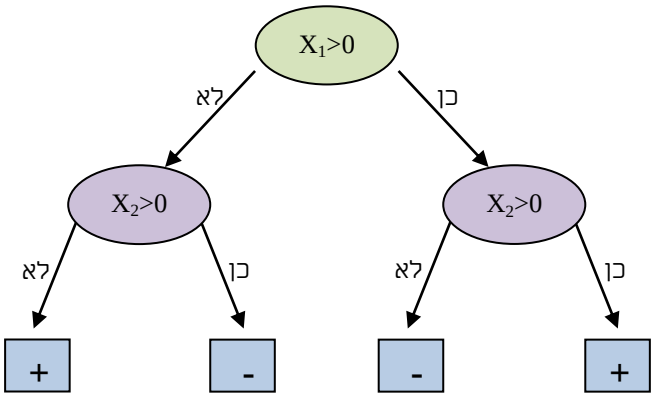
* ז. אופטימליות מול חמדנות

בהנתן קבוצת דוגמאות מתוייגות, האם ניתן למצוא בזמן סביר עץ בעל עומק קטן ככל האפשר המתייג נכון את על הקבוצה? רון ריבסט ולאורנט הייפיל הראו בשנת 1976 כי זו בעיה NP-Complete, השקולה במורכבותה לבעיות רבות אחרות הידועות כקשות לפתרון יעיל.
כתוצאה מכך אלגוותמים לבניית עצי החלטה הם חמדניים, כפי שראינו. בכל שלב האלגוריתם בוחר מאפיין על פי מדד איכות, שערכו אי-שלילי, וערך גבוה מאפיין איכות טובה יותר. האם יש צורך לעצור את אלגוריתם הלמידה כאשר ערך המדד הוא מינימלי 0 לכל מאפיין.

נבחן את בעיית ה-xor שבה מתקיים $y = sign(x_1 x_2)$. נניח כי לפנינו ארבע דוגמאות על פי הטבלה הבאה:

input	label
(1,1)	1
(-1,-1)	1
(-1,1)	-1
(1,-1)	-1

ניתן לראות כי על מאפיין לבדו אינו מועיל ובעל מדד עם ערך 0, ולכן אלגוריתם חמדני שיפסיק להוסיף קודקודים כאשר ערך המדד 0 יחזיר עץ ריק. מצד שני קיים עץ בעומק 2 כלדקמן:



8.3 הרחבות

א. מאפיינים רציפים

נניח כי וקטור המאפיינים $x = (x_1, \ldots, x_d)$ כולל רכיבים x_i בעלי ערכים רציפים. במקרה זה, המבחן המקובל לגבי x_i הינו מהצורה $x_i \le t_i$. לפיכך, לבחירת המאפיין כל צומת יש להוסיף את בחירת ערך הסף t_i.

עבור כל מבחן $A = \{x_i \le t_i\}$ ניתן להגדיר את תוספת המידע באופן הרגיל:

$$Q(S \mid x_i, t_i) = Q(S \mid A)$$

השלב הבא הוא מכסימיזציה על הסף t_i:

$$Q(S \mid x_i, t_i^*) = \max_{t_i} Q(S \mid x_i, t_i^*)$$

ולאחר מכן בחירת המאפיין x_i שעבורו מדד זה הינו מכסימלי.

ב. ערכי יציאה רציפים (בעיית רגרסיה)

כאשר ערכי המוצא $y = f(x)$ הינם רציפים, נדרשים מספר שינויים בהגדרות הקודמות. ראשית, ערכי המוצא בעלי העץ (אשר בבעיית סיווג נקבעו לפי הרוב) ייקבעו עתה לפי הממוצע: בעלה ℓ שאליו הגענו עם קבוצת דוגמאות $S_\ell = \{x_k, d_k\}_{k=1}^{N_\ell}$, ערך המוצא ייקבע לפי

$$y_\ell = ave\{d_k\}_{k=1}^{N_\ell} = \frac{1}{N_\ell}\sum_{k=1}^{N_\ell} d_k$$

שנית, מדדי חוסר-האחידות $Q(S)$ שהגדרנו לבעיית הסיווג אינם מתאימים פה. ההגדרה המקובלת פה הינה הסטייה הריבועית מהממוצע:

$$Q(S) = \frac{1}{N}\sum_{k=1}^{N}(d_k - \hat{d})^2$$

כאשר $S = \{x_k, d_k\}_{k=1}^{N}$ ו- $\hat{d} = ave\{d_k\}_{k=1}^{N}$. בכל שאר ההגדרות אין שינוי.

* 8.4 התאמת יתר וגיזום

גם בתהליך הבניה של עצי החלטה קיימת הבעיה של התאמת יתר, כלומר התאמה מדויקת "מדי" לדוגמאות שמביאה להגדלת השגיאה בסיווג תבניות קלט חדשות.

ניתן לגשת לבעיה זו בשתי צורות:

1. הגבלת אורך העץ בזמן בנייתו: נתן למשל להגביל מראש את האורך המכסימלי המותר של ערוץ החלטה כלשהו. ניתן גם לקבוע סף תחתון ל"תוספת המידע" הנדרשת להוספת מאפיין בצומת כלשהו בעץ.
2. גיזום (pruning) – קיצור העץ לאחר גמר בנייתו.

גישת הגיזום התבררה כמוצלחת יותר בפועל, ולפעולת הגיזום השפעה רבה על הביצועים המתקבלים.

גישה מקובלת לבקרת תהליך הגיזום הינה באמצעות קבוצת ביקורת, הנפרדת מקבוצת הלימוד, ומאפשרת בדיקת השפעת שינויים אפשריים על שגיאת הסיווג. מקובל להקצות כ- 1/3 מכלל הדוגמאות לקבוצת הביקורת. אפשרות עדיפה כאשר מספר הדוגמאות מוגבל הינה שימוש באימות-צולב.

נתאר ביתר פרוט שיטה מסוימת לביצוע הגיזום. נסמן את העץ הראשוני (המכסימאלי) על ידי T_0 T_0. גיזום מתבצע על ידי הורדת מספר כלשהו של תת-עצים המתחילים בצמתים פנימיים של T (צמתים פנימיים אלה הופכים לעלים בעץ החדש T). לכל עץ גזום T נסמן ב- L_T את קבוצת העלים, ולכל עלה ℓ תהי S_ℓ קבוצת הדוגמאות בעלה זה ו- N_ℓ מספרן. נגדיר עתה מדד ביצועים עם קנס על גודל העץ כלהלן:

$$C_\alpha(T) = \sum_{\ell=1}^{|L_T|} N_\ell Q(S_\ell) + \alpha \,|\, L_T \,|$$

עבור כל פרמטר α נתון, ניתן למצוא את העץ הגזום $T = T_\alpha$ המביא למינימום קריטריון זה. שיטה יעילה לכך הינה *גיזום החוליה החלשה*: בכל שלב גוזמים את הצומת הפנימי אשר מביא לעליה הקטנה ביותר בגודל $\sum_\ell N_\ell Q(S_\ell)$, וממשיכים עד לקבלת עץ חד-צמתי. ניתן להראות כי לכל α, T_α נמצא בקבוצת העצים שהתקבלו באופן זה. בחירת α מתבצעת בעזרת סדרת הביקורת.