תרגול 8 : עצי החלטה

עצי החלטה – תקציר התאוריה

מטרת חלק זה היא להדגים בנייה של עץ החלטה, כאשר בחירת המאפיין המיטבי נעשית עפ"י קריטריון אנטרופיה.

יהי S אוסף של N דוגמאות מסווגות $S=\left\{x_k,d_k\right\}_{k=1}^{N}$, כך ש- $d_k\in\{c_1,\ldots,c_C\}$.

השכיחות היחסית (או "הפילוג האמפירי") של כל אחד מהסיווגים האפשריים (c_j) בקבוצת הדוגמאות נתונה ע"י:

$$\hat{p}_j=\frac{1}{N}\sum_{k=1}^{N}I\{d_k=c_j\},\quad j=1,\ldots,C$$

מדדים לחוסר אחידות של S:

1. שגיאת הסיווג: $Q(S)=1-\max\limits_{j\in\{1,\ldots,C\}}\hat{p}_j$
2. אינדקס Gini: $Q(S)=\sum_j\hat{p}_j(1-\hat{p}_j)$
3. אנטרופיה: $Q(S)=H(S)=\sum_j\hat{p}_j\log_2\frac{1}{\hat{p}_j}=-\sum_j\hat{p}_j\log_2(\hat{p}_j)$

תכונות של $Q(S)$:

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

תוספת המידע של מאפיין:

נניח כי מאפיין A כלשהו מחלק את S למספר תת-קבוצות. נסמן תת-קבוצות אלו על ידי $\left\{S_m,\ m\in 1,2,\ldots,M\right\}$, כאשר M הינו אוסף הערכים האפשריים של A. <u>מדד חוסר-האחידות המשוקלל</u> עבור האוסף $\{S_m\}$ יוגדר עתה על ידי:

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

כאשר $Q\left(S_m\right)$ הוא מדד לחוסר האחידות של תת-הקבוצה S_m.
מדד הטיב של המאפיין 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 המאפיין A.

המאפיין A שנבחר הוא (כעיקרון) זה שעבורו השיפור $\Delta Q(S \mid A)$ הינו מקסימלי, כלומר $Q(S \mid A)$ מינימלי.

שאלה 1 – בניית עץ החלטה

בנה עץ החלטה המבוסס על קריטריון האנטרופיה, אשר בהינתן נתוני צבע שער, גובה, משקל, משתמש בקרם הגנה, קובע האם עתיד האדם להכוות מהשמש היוקדת.

סט דוגמאות הלימוד לצורך בניית העץ מוצג בטבלה הבאה:

Independent Attributes / Condition Attributes — Dependent Attributes / Decision Attributes

Name	**Hair**	**Height**	**Weight**	**Lotion**	**Result**
Sarah	blonde	average	light	no	sunburned (positive)
Dana	blonde	tall	average	yes	none (negative)
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none