

תורת האינפורמציה

נקודת מבט

כמות האינפורמציה המקודדת במשתנה מקרי X עם התפלגות ידועה.

שאלה: כאשר נדרוש תוצאה של משתנה - מהי כמות האינפורמציה (בביטים) שמועברת בתצפית, ונרצה לדעת מה תוחלת כמות המידע עבור המשתנה.

נסמן

אנטרופיה - $H[X]$ - תוחלת כמות האינפורמציה שצריך בשביל לייצג את המשתנה

ניזכר

כמו שלמדנו¹ לגבי קוד הופמן - בקוד אופטימלי נשתמש ב $\log p(x)$ ביטים כדי לקודד מאורע x

דוגמה

משתנה מקרי עם 4 ערכים אפשריים A, B, C, D : $p(x) : \langle \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \rangle$. נקודד:

x	=	A	B	C	D
$p(x)$	=	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
code	=	00	01	10	11

$$H[X] = 2$$

לעומת זאת, אם $p(x) : \langle \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} \rangle$ נקודד:

x	=	A	B	C	D
$p(x)$	=	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
code	=	0	10	110	111

$$H[X] = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}$$

באופן כללי

$$H[X] = - \sum_{x \in X} p(x) \cdot \log p(x)$$

$$0 \leq H[X] \leq \log |X|$$

האנטרופיה מקסימלית עבור התפלגות אחידה, ומתאפסת כאשר X קבוע.

¹לא בקורס הזה

Perplexity של משתנה מקרי X

נגדיר perplexity של משתנה מקרי X :

$$2^{H[X]}$$

אינטואיטיבית: משמעות המושג היא שרמת האינפורמציה (אי הוודאות) במשתנה שקולה למשתנה מקרי יוניפורמי עם אותו ערך perplexity.

למשתנה ברנולי

$$H[x] = -p \log p - (1-p) \log(1-p)$$

Cross Entropy ("אנטרופיה צולבת")

נקודת מבט: מספר הביטים הנדרש לקודד מאורעות שנדגמו לפי התפלגות p , כאשר משתמשים בקוד אופטימלי להתפלגות q .

(לדוגמה: $p = q$ = ההתפלגות האמיתית, q = המודל שיש לנו)

$$H[p||q] = - \sum_{x \in X} p(x) \log q(x)$$

נשים \heartsuit : $H[p||q]$ אינו סימטרי

מתקיים (תיכף נוכיח):

$$H[p||q] \geq H[p]$$

משמעות: תמיד נפסיד ביטים אם נקודד את p עם $q \neq p$.

הגדרה - Relative Entropy (אנטרופיה יחסית) $\equiv (D_{KL})$ KL-Divergence

אנטרופיה יחסית תוגדר תוגדר כמספר הביטים שמאבדים כאשר ממדלים את p ע"י q :

$$\begin{aligned} D_{KL}[p||q] &= H[p||q] - H[p] = - \sum_x p(x) \log q(x) - \left[- \sum_x p(x) \log p(x) \right] = \\ &= - \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)} \end{aligned}$$

זכור: • לא מוגדר אם קיים x שעבורו $q(x) = 0$ ו $p(x) \neq 0$

• שוויון כאשר $p = q$ $\iff D_{KL}(p||q) \geq 0$

• D_{KL} לא סימטרי

שימוש ב: D_{KL}

נשתמש הרבה כמדד לסטייה בין התפלגויות. בד"כ מעוניינים לצמצם אותו כסטייה בין התפלגות משוערכת לאיזושהי התפלגות "אופטימלית"

מדד סימטרי לסטייה בין התפלגויות - Jensen-Shanon Divergence

$$J(p||q) = \frac{1}{2} \left[D \left(p \left\| \frac{p+q}{2} \right. \right) + D \left(q \left\| \frac{p+q}{2} \right. \right) \right]$$

תמיד מוגדר

מדד אסימטרי שמכליל את J :

נותן משקול שונה (כפרמטר) לשני המרחקים בנוסחת J . נקרא Skew-Divergence.

קירובי אמפירי ל Cross Entropy

משמעות נוספת: קירוב לחסם של האנטרופיה של משתנה

נסתכל על משתנה שמייצר סדרות - שפה L .

בשלב ראשון - נניח שכל הסדרות באורך קבוע n . נסמן כל סדרה: $W_1^n \ni w_1, \dots, w_n \doteq w_1^n$.

$$H[W_1^n] = - \sum_{w_1^n \in W_1^n} p(w_1^n) \log p(w_1^n)$$

המשמעות: מספר הביטים בממוצע לדווח על תצפית של סדרה מהמשתנה.

הגדרה - Entropy Rate

$$\frac{1}{n} H[W_1^n]$$

עבור שפה L כהליך סטטיסטי(אינסופי) נגדיר Entropy Rate:

$$H[L] = \lim_{n \rightarrow \infty} \frac{1}{n} H[W_1^n]$$

משפט Shannon-McMillen-Breiman

עבור שפה L שמקיימת תכונות:

1. stationary - אין תלות בנקודת הזמן

2. ergodic - אין תלות ב"עבר הרחוק". התלות בעבר היא חסומה

מתקיים:

$$H[L] = \lim_{n \rightarrow \infty} -\frac{1}{n} \log p(w_1^n)$$

האינטואיציה: סדרה באורך ששואף לאינסוף מייצגת את המודל, ואין צורך לחשב תוחלת על כל הסדרות האפשריות. השימוש הפרקטי: מספיק לשערך את $p(w_1^n)$ על סדרה אחת, מספיק ארוכה. כיוון ש p אינו ידוע, ויש לנו רק מודל משוערך q , נרצה לאמוד את cross entropy rate $H[p||q]$:

$$H[p||q] = \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{w_1^n \in W_1^n} p(w_1^n) \log q(w_1^n)$$

לפי המשפט הנ"ל, מיושם לגבי $H[p||q]$:

$$H[p||q] = \lim_{n \rightarrow \infty} -\frac{1}{n} \log q(w_1^n)$$

השימוש המעשי: ניקח סדרה אחת מספיק ארוכה, נחשב $\log q(w_1^n)$ לפי המודל המשוערך שלנו q , וכך נקבל אומדן ל $H[p||q]$.

קיבלנו: שככל שהמודל q נותן הסתברות גבוהה יותר לתצפית ארוכה שנוצרה ע"י p , המשמעות היא ש $H[p||q]$ קטן יותר - כלומר המודל מייצג טוב יותר את התצפית במונחי אובדן אינפורמציה.