מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 17-01-2016

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

תורת האינפורמציה

נקודת מבט

כמות האינפורמציה המקודדת במשתנה מקרי X עם התפלגות ידועה.

נשאל: 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 שלמדנו[1] 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: $p(x):\left\langle \frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right\rangle$. 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 $p(x):\left\langle \frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right\rangle$ 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.

[1] לא בקורס הזה

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 17-01-2016

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

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 = ההתפגלות האמיתית, q = המודל שיש לנו)

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

נשים ♡: $H[p\|q]$ 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) $\equiv$ KL-Divergence(D_{KL})

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

$$D_{KL}[p\|q] = H[p\|q] - H[p] = -\sum_x p(x)\log q(x) - \left[-\sum_x p(x)\log p(x)\right] =$$

$$= -\sum_x p(x)\log\frac{p(x)}{q(x)} = E_{p(x)}\log\frac{p(x)}{q(x)}$$

כזכור:

- לא מוגדר אם קיים x שעבורו $q(x)=0$ ו$p(x)\neq 0$
- שוויון כאשר $p=q \Longleftarrow D_{KL}(p\|q)\geq 0$
- D_{KL} לא סימטרי

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 17-01-2016

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

שימוש ב:D_{KL}

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

מדד סימטרי לסטיה בין התפלגויות - Jensen-Shanon Divergence

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

תמיד מוגדר

מדד אסימטריה שמכליל את J:

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

קירובי אמפירי לCross 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: $W_1^n \ni w_1,\dots w_n \triangleq w_1^n$.

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

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

הגדרה - Entropy Rate

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

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

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

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:

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

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 17-01-2016

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

האינטואיציה: סדרה באורך ששואף לאינסוף מייצגת את המודל, ואין צורך לחשב תוחלת על כל הסדרות האפשריות.

השימוש הפרקטי: מספיק לשערך את $p\left(w_1^n\right)$ על סדרה אחת, מספיק ארוכה.

כיוון שp אינו ידוע, ויש לנו רק מודל משוערך q, נרצה לאמוד את ה$H\left[p\|q\right]$ cross entropy rate:

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

לפי המשפט הנ"ל, מיושם לגבי $H\left[p\|q\right]$:

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

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

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