מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 2015-11-08

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

התפלגות מולטינומית Multinomial

משתנה X שמקבל m ערכים שונים $x_1, \dots x_i, \dots x_m$, כל ערך x_i בהסתברות p_i, $\sum_{i=1}^{m} p_i = 1$.

נוסחת הסתברות למולטינום:

עבור n דגימות ממולטינום, ההסתברות לקבלת k_i דגימות מכל ערך x_i נקבעת לפי הנוסחה:

$$P_n\left(k_1, k_2, \dots k_m\right) = \frac{n!}{k_1! \cdots k_m!} \cdot p_1^{k_1} \cdots p_m^{k_m}$$

נתעניין בכל פעם בערך יחיד של המשתנה, ניתן להסתכל עליו כמשתנה בינומי כשהערך המבוקש נחשב הצלחה, אחרת כישלון.

אומדן ML:

$$p_i = p\left(x_i\right) = \frac{k_i}{n}$$

לפי n דגימות

תוחלת מספר הפעמים לקבלת ערך x_i בn דגימות:

כמו בבינום: $n \cdot p_i$

מודל שפה - Language Model

התופעה שרוצים למדל:

יצירה של רצפים של סימנים מתוך אוסף דיסקרטי של סימנים אפשריים.

מבט גנרטיבי:

יש מקור שמייצר את רצף הסימנים לפי מודל הסתברותי כלשהו.

מטרת מודל שפה הסתברותי:

בהינתן רצף סימנים, לחשב את ההסתברות ליצירת הרצף

לדוגמה:

במערכת לזיהוי דיבור, נרצה מודל שפה שיידע לאמוד את ההתסברות של כל משפט אפשרי, ובהתאם המערכת לעדיף לייצר משפטים סבירים יותר.

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 2015-11-08

פיתוח מודל שפה הסתברותי

בהינתן דגימה של סדרת סימנים בשפה: $s = w_1, \dots w_n$, נרצה למדל את $P\left(w_1, \dots w_n\right)$
צעד ראשון:

$$P\left(w_1, \dots w_n\right) = P\left(\text{length}\left(s\right) = n\right) \cdot p\left(s | \text{length}\left(s\right) = n\right)$$

נשים ♡: $\sum_{n=1}^{\infty} P\left(\text{length}\left(s\right) = n\right) = 1$

עבור n מקובע:

$$\sum_{\substack{w_1, \dots w_n \\ \forall_i w_i \in W}} P\left(w_1, \dots w_n | \text{length} = n\right) = 1$$

(כאשר W - סט כל המילים האפשריות)

מודול הסתברות הסדרה - נניח ש$n = |s|$

נשים לב שמספר המשפטים האפשריים מאוד מאוד גדול($|W|^n$) - הרבה יותר גדול ממספר הדגימות שיש לנו. לכן נרצה לפרק את המשפט.

$$P\left(w_1, \dots w_n\right) = P\left(w_1\right) \cdot P\left(w_2 | w_1\right) \cdot P\left(w_2 | w_1, w_2\right) \cdots P\left(w_n \middle| w_1^{n-1}\right)$$

כדי לקבל מודל שניתן לשערך את ההסתברויות שלו, נניח הנחות אי תלות. למשל - הסתברויות המילים בסדרה אינן תלויות אחת בשנייה:

$$\overset{\text{assuming}}{\underset{\text{independence}}{}} = P\left(w_1\right) \cdot P\left(w_2\right) \cdots P\left(w_n\right) = \prod_{i=1}^{n} P\left(w_i\right)$$

קיבלנו מודל מולטינומי, עם $|W|$ ערכים אפשריים.

המודל הגנרטיבי: בכל מקום i, מגרילים מילה w_i בהסתברות $P\left(w_i\right)$: $P_{ML}\left(w\right) = \frac{\text{freq}(w)}{N}$(כאשר N - גודל המדגם)

דוגמאות לשימושים

- סיווג מאמרים - למשל אם מקבלים מסמך d, וידוע לנו ההסתברות של מילים $w \in W$ להימצא במאמר ספורט $P^S\left(w\right)$ ומה ההסתברות שלה להימצא במאמר משפטי $P^L\left(w\right)$, ואז אם $P^L\left(d\right) > P^S\left(d\right)$ נסווג את המאמר בתור מאמר משפטי ולא בתור מאמר ספורט.
- זיהוי דיבור - אם יש שתי היפותזות לגבי מה שהמשתמש אמר, נעדיף את היפותזה א' על פני ב' אם $P^*\left(\text{א}\right) > P^*\left(\text{ב}\right)$

המודל המולטינומי הוא לא האפשרות היחידה. נראה כעת מודל אחר(נראה אותו תחת הדוגמה של סיווג מסמכים):

דוגמה למודל עבור קבוצות(סט) מילים

נשים ♡: בקבוצה אין חזרות ואין חשיבות לסדר. במודל המולטינומי, אמנם כאשר מניחים אי תלות אין חשיבות לסדר - אבל עדיין יש חשיבות לחזרות.

Multiple Bernouli

כדי למדל קבוצת סימנים שנצפתה(למשל - קבוצת המילים שהופיעה במסמך לצורך סיווג טקסטים), נתייחס ליצירת הקבוצה כדגימה ממספר משתני ברנולי: לכל מילה אפשרית בשפה $w \in W$ נגדיר:

$P(w)$ ההסתברות שהמילה תופיע במסמך אקראי

ובהתאם - $1 - P(w)$ - ההסתברות שלא תופיע

תהליך היצירה: לכל מילה במילון, נגריל לפי הסתברות $P(w)$ באם לכלול אותה במסמך.

נסמן בd מסמך אקראי: $d \subseteq W$:

$$P(d) = \prod_{w \in d} P(w) \cdot \prod_{w \notin d} (1 - p(w))$$

נשים ♡: אפשר להסתכל על תת קבוצה d כוקטור בינרי באורך $|W|$

אומדן $P(w)$:

$$P(w) = \frac{\text{number of documents containing } w}{\text{number of ducuments in the sampling}}$$

נשווה לעומת המודל המולטינומי, שבו

$$P(w) = \frac{\text{number of occurences of } w \text{ in all the texts}}{\text{number of occurences of all the words in all the texts}}$$

ראוי לציין

בדרך כלל כשרוצים למדל שפה טבעית, עדיף להשתמש במודל מולטינומי כי יש חשיבות לשכיחות של המילים.

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 2015-11-08

הקדמה לשיעור הבא - מידול רצפי סימנים ע"י מודל n-gram

אינטואיצה: מניחים שההסתברות של מילה מושפעת מהמילים לפניה - אבל לא יותר מדי, מטעמי גודל מדגם ומטעמי חישוביות.

טעם נוסף זה שככל שהולכים יותר אחורה מאבדים הקשר - אבל בדרך כלל מגיעים למגבלה הטכנית הרבה לפני שזה קורה.

המודל המולטינומי:

$$P\left(w_1,\ldots w_n\right)=P\left(w_1\right)P\left(w_2|w_1\right)\cdots P\left(w_n|w_1,\ldots w_{n-1}\right)$$

ולאחר הנחות אי תלות מקבלים מודל יוניגרם:

$$P\left(w_1\right)P\left(w_2\right)\cdots P\left(w_n\right)$$

מודל ביגרם:

$$P\left(w_1\right)P\left(w_2|w_1\right)P\left(w_3|w_2\right)\cdots P\left(w_n|w_{n-1}\right)$$

מודל טריגרם:

$$P\left(w_1\right)P\left(w_2|w_1\right)P\left(w_3|w_1,w_2\right)\cdots P\left(w_n|w_{n-2},w_{n-1}\right)$$

ההנחות: יוניגרם $P\left(w_i|w_1,\ldots w_{i-1}\right)=P\left(w_i\right)$
ביגרם $P\left(w_i|w_1,\ldots w_{i-1}\right)=P\left(w_i|w_{i-1}\right)$
טריגרם $P\left(w_i|w_1,\ldots w_{i-1}\right)=P\left(w_i|w_{i-2},w_{i-1}\right)$
n-גרם $P\left(w_i|w_1,\ldots w_{i-1}\right)=P\left(w_i|w_{i-n+1},\ldots w_{i-1}\right)$