

שיערוך Back off

הרעיון: אם לא מוצאים ב-n-גרם, נחפש ב-1-n-גרם

$$P_B = (w'|w) = \begin{cases} P_d(w'|w) & C(w, w') > 0 \\ \alpha(w) P_d(w') & \text{otherwise} \end{cases}$$

$$P \left(w_n \left| \overbrace{w_1, \dots, w_{n-1}}^{w_1^{n-1}} \right. \right) : \text{נשמך } w_1, \dots, w_n : n\text{-gram}$$

למשל: הרעיון ל-3-gram:

- אם אין סטטיסטיקה (שכיחות 0) ל- w_1^3 , נעשה backoff ל- w_2^3
- אם אין סטטיסטיקה ל- w_2^3 , נעשה backoff ל- w_3

נוסחה כללית ל- P_B ל-n-gram:

$$P_B(w_n | w_1^{n-1}) = \begin{cases} P_d(w_n | w_1^{n-1}) & C(w_1^n) > 0 \\ \alpha(w_1^{n-1}) \cdot P_B(w_n | w_2^{n-1}) & \text{otherwise} \end{cases}$$

עבור יוניגרם: $P_B(w) = P_d(w)$, כי אין למה לעשות backoff...

מה זה α ?

עבור מילה מתנה בודדת:

$$\alpha(w) = \frac{1 - \sum_{w': C(w, w') > 0} P_d(w'|w)}{1 - \sum_{w': C(w, w') > 0} P_d(w')}$$

ועבור סדרת מתנה של $n-1$ מילים:

$$\alpha(w_1^{n-1}) = \frac{1 - \sum_{w_n: C(w_1^n) > 0} P_d(w_n | w_1^{n-1})}{1 - \sum_{w_n: C(w_1^n) > 0} P_d(w_n | w_2^{n-1})}$$

במילים: הסיכוי למילה לא ידועה אחרי w_1^{n-1} חלקי הסיכוי למילה לא ידועה אחרי w_2^{n-1} .

הרעיון הוא שרק אם אין לנו מידע אנחנו עושים backoff.

גישות אינטרפולציה לאומדן

יש לנו כמה דרכים שונות לשערך התסברויות, ואנחנו רוצים לעשות אינטרפולציה בין התוצאות שלהן. הרעיון: נשלב בצורה ממושקלת מספר אומדנים.

דוגמאות: • שילוב בין מודל ביגרם למודל נאיבי ($0 < \lambda < 1$):

$$P_{\text{int}}(w'|w) = \lambda \cdot P_{ML}(w'|w) + (1 - \lambda) \cdot P(w')$$

• מודל יוניגרם לטקסטים ברפואה: נשלב מודל שנלמד מטקסטים ברפואה למודל מטקסטים כלליים:

$$P(w) = \lambda \cdot P_{\text{med}} + (1 - \lambda) \cdot P_{\text{gen}}(w)$$

• באופן דומה לשקלול מספר מודלים:

$$P(w_3|w_1, w_2) = \lambda_1 \cdot P(w_3|w_1, w_2) + \lambda_2 \cdot P(w_3|w_2) + \lambda_3 \cdot P(w_3)$$

$$\sum_{i=1}^3 \lambda_i = 1 \text{ ש } \lambda_i = 1$$

איך מכיילים?

- במקרה פשוט, של מעט פרמטרים λ_i , ניתן לבצע כיול אמפירי ע"י ניסוי מספר ערכים.
- נשים \heartsuit : במודל n -gram ניתן לקבוע צירוף מקדמים שונה לכל התפלגות, כלומר עבור כל מאורע מתנה. בפרט: נצפה לערכי λ אופטימליים גבוהים יותר מאורעות מתנים שכיחים יותר. ניתן לקבץ מאורעות מתנים לפי שכיחויות ולכיל מקדמים ביחד לכל קבוצה.
- יש מודלים של אינטרפולציה שעבורם קיימים אלגוריתמים יעילים לכיול המקדמים.

בפרט: ל n -gram יש אלגוריתם ME

משתנים חבויים - Hidden Variables

לפעמים נרצה לחשב אספקטים נוספים של התופעה, שאנחנו לא צופים בהם, כדי למדל יותר טוב את התופעה.

סכימה כללית

נתונה תופעה נצפית Y , בד"כ מורכבת $y = (y_1, \dots, y_n)$.

דוגמה: סקירת מילים/ידיעה/מסמך

מניחים שקיימת התפלגות משותפת של Y עם משתנה חבוי נוסף X

דוגמה: הסדק במערכת שבו נכתבה הידיעה

בד"כ נתעניין ב- X דיסקרטי שמקבל מספר סופי של ערכים "אטומיים".

בהליך הגנרטיבי שנניח: כדי לייצר תצפית $y \in Y$ קודם נבחר $x \in X$ לפי $p(x)$, ואח"כ נייצר y לפי $p(y|X=x)$.

שימושים במודלים חבויים

כללית: גם כשאין שימוש לערכי X , מידול בעזרת X יכול לשפר את מידול Y .

במקרים אחרים יש עניין יישומי בערכי X :

- בסיטואציות אופייניות של למידה/סיווג:

1. למידה/סיווג מבוקר/מפוקח (Supervised):

ערכי X ידועים, ויש מדגם אימון שבו לכל $y \in Y$ שנצפה ידוע(מתוייג) מה ערך $x \in X$ המתאים לאותו y .

- ממדגם כזה נשערך את ההתפלגות המשותפת.

- יישום: בהינתן y , לשערך את $p(x|y)$. ובפרט: לסווג את x המתאים.

2. למידה לא מפוקחת (Unsupervised):

לא ידועה מראש זהות ערכי X , ואין מדגם מתוייג.

המטרה: לקבל(קלסטרינג) את התצפיות של y לפי ערכי $x \in X$ שכנראה ייצרו אותם.

מטרת הלמידה: שערך θ , אוסף הפרמטרים של $p(x, y; \theta)$

מודל חבוי - עירוב היסטוגרמות Mixture of Histograms

(זה בעצם עירוב של מולטינומים)

נמיש: כמודל ליצירת מסמכים (Y) לפי נושאים (X).

נתייחס למסמך כהיסטוגרמת שכיחויות של המילים שמופיעות בו(זה המודל המולטינומי שלמדנו ליוניגרם).

נסמן: המילים נלקחות ממילון בגודל v : w_1, \dots, w_v

מדגם מסמכים: $y = y_1, \dots, y_N$ (N מסמכים במדגם)

נניח אי תלות בין המסמכים.

ייצוג מסמך y_t : כהיסטוגרמה של v שכיחויות $(n_{t_1}, n_{t_2}, \dots, n_{t_v})$ כאשר n_{t_k} היא שכיחות המילה w_k במסמך y_t

לפשטות: נניח אורך מסמכים קבוע n : $\forall_t \sum_{k=1}^v n_{t_k} = n$

מודל עם משתנה חבוי

נניח קיום:

- מ"מ נוסף X עם ערכים $x_1, \dots, x_{|X|}$ (בדוגמה המסמכים: מייצג קטגוריה, או קלסטר)
- התפלגות משותפת $p^1(x, y; \theta)$

תהליך גנרטיבי ליצירת מסמך

1. נבחר $x \in X$ (קטגוריה) לפי התפלגות $p(x) = P(X = x)$

2. נקבע את המילים y לפי התפלגות $p(y_t | X = x)$

בפרט: נניח עבור $p(y_t | x)$ מודל מולטינומי נפרד לכל $x \in X$

בהתאם - הפרמטרים θ של המודל

$$\forall x_i \in X p(x_i)$$

$$\forall x_i \in X \forall w \in w_1, \dots, w_v p(w_k | x_i)$$

לפי המודל המולטינומי:

$$p(y_t | x_i) = \prod_{k=1}^v p(w_k | x_i)^{n_{t_k}}$$

נתעניין ב3 שאלות:

1. $p(y_t; \theta)$ - הסתברות יצירת תצפית
2. $p(x_i | y_t; \theta)$ - הסתברות הסיווי/שיוך לקלסטר
3. אומדן $\hat{\theta}_{ML}$ - נסיון ערכי ML לפרמטרים