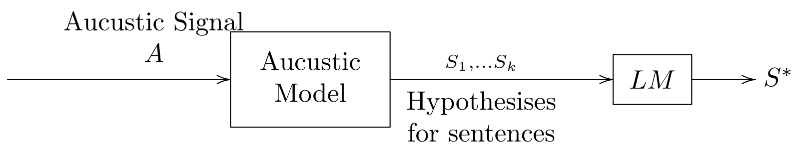
מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 2015-11-15

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

דוגמאות לשימוש במודל שפה במערכות יישומים

דוגמה 1 - זיהוי דיבור



$$P\left(S|A\right)=\frac{P\left(A|S\right)P\left(S\right)}{P\left(A\right)}$$

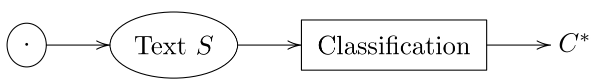
$$S^{*}=\mathrm{argmax}_{S}P\left(S|A\right)=\mathrm{argmax}_{S}=\mathrm{argmax}_{S}\underbrace{P\left(A|S\right)}_{\text{Aucustic Model}}\cdot\underbrace{P_{LM}\left(S\right)}_{LM}$$

כנ"ל לתרגום: $A\equiv$שפת המקור

$P\left(A|S\right)\equiv$מודל תרגום משפה A לשפה S

צריך לבחור מודל יחיד - תצפית לא וודאית ממספר היפותזות

דוגמה 2 - סיווג טסקטים



החלקות - Smoothing

מודל n-gram: קירוב מרקובי

הנחת קירוב מרקובי - ההסתברות למאורע תלוייה רק במספר מוגבל של מאורעות אחורה:

$$p\left(w_i \middle| w_1^{i-1}\right) = p\left(w_i \middle| w_{i-n+1}^{i-1}\right)$$

במודל n-gram יש פרמטר לכל n-יה סדורה של מילים שמגדיר את ההסתברות המותנית המתאימה(של מילה בהינתן n-1 המילים הקודמות)

לדוגמה: ב3-gram מסתכלים על שתי מילים אחורה - כלומר במודל יש פרמטר לכל שלשת מילים - לדוגמה עבור הטקסט I ate a peach יש פרמטרים כמו $p\left(peach|ate,a\right) \qquad p\left(a|I,ate\right)$

נסמן: V - לקסיקון. מספר הפרמטרים:$|V|^n$

הערות מימוש:
- בפועל נשמור פרמטרים רק לn-יות שנצפו במדגם, וזה תת-לינארי בגודל המדגם.
- יוריסטיקה - בד"כ כאשר השפה מאוד גדולה, נתעלם מרצפים שראינו רק פעם אחת.

נשים ♡: המודל מגדיר התפלגות מולטינומית נפרדת מעל V לכל סדרת מילים מתנה באורך $n-1$ ($n-1$-יה) - סה"כ $|V|^{n-1}$ התפלגויות שונות)

אומדן MLE לפרמטרים

$$p_{MLE}\left(w_i \middle| w_{i-n+1}^{i-1}\right) = \frac{\#\left(w_{i-n+1}^{i}\right)}{\#\left(w_{i-n+1}^{i-1}\right)}$$

(# - מונה)

לדוגמה - עבור I ate a peach

$$p_{MLE}\left(peach|ate,a\right) = \frac{\#\left(\text{times seen "ate a peach"}\right)}{\#\left(\text{times seen "ate a *"}\right)}$$

נשים ♡ אם הn-יה לא נצפתה, אזי נשערך לה הסתברות 0 (נניח בשלב זה שהw_{i-n+1}^{i-1} נצפו). מצב זה יוצר בעיה כאשר יתכנו מאורעות יחסית נדירים, שלא נצפו במדגם, אבל ההתסברות האמיתית שלהן עדיין חיובית.

בגלל הצורה הכפלית של מודל n-gram מספיק שלא נצפה בn-יה אחת ונקבל $P\left(S\right)=0$

הפתרון לבעיה הזו - החלקה

החלקה - Discounting

במרחב מסוג מודל שפה, שבו יש הרבה מאורעות נדירים שאפשריים אך לא ייצפו במדגם נתון, נרצה לבצע Discounting - לתת אומדן חיובי למאורעות שלא נצפו, ולמאורעות שנצפו לתת אומדן נמוך מאומדן הMLE.

נזכור: באופן יחסי - האומדנים אמינים יותר למונים גבוהים, ועבורים יהיה נכון לעשות הפחתה יחסית קטנה יותר לאומדן MLE

מודל החלקה ראשון - Lidstone

נקודת מבט של הוספת קבוע λ למונה של כל מאורע אפשרי, ונרמול בהתאם:

$$p_{Lid}\left(x\right)=\frac{C\left(x\right)+\lambda}{\left|S\right|+\lambda\left|X\right|}$$

(C - מונה)

- עבור $\lambda=1$: נקרא החלקת Laplace
- כללית: λ הוא פרמטר של השיטה

ההצדקה: אומדן לידסטון מוצדק כאומדן בייסיאני אם מניחים הסתברות א-פריורית אחידה לכל המאורעות

מתקיים: האומדן הוא אינטרפולציה לינארית בין התפלגות אחידה לאומדן MLE:

$$p_{Lid}\left(x\right)=\mu\frac{C\left(x\right)}{\left|S\right|}+\left(1-\mu\right)\frac{1}{\left|X\right|}$$

עבור

$$\mu=\frac{\left|S\right|}{\left|S\right|+\lambda\left|X\right|}$$

μ הוא הפרופורציה של $\left|S\right|$ מתוך "המדגם המוגדל"

שימושיות: שיטה מאוד פשוטה ונפוצה לשימוש

מוצדקת: כשיש prior אחיד

מצד שני: תיתן אומדנים מוטים במקרים אחרים - שיכולים לגרום נזק(בפועל - כתלות ברגישות המערכת הכוללת להחלקה)

בפרט - נראה בעייתיות של מודל Lindstone ליישומים שבהם נרצה התנהגות Discounting

ובפרט נראה שהחלקת Lindstone מגדילה את האומדן למאורעות מסויימים **שנצפו** ביחס לאומדן MLE.
כיוון שמתנהגם כאינטרפולציה לינארית - ערך P_{lid} יהיה "באמצע"(מוטה ע"י מקדם האינטרפולציה) בין P_{MLE} להתפלגות יוניפורמית.
נשים ♡: אם $C\left(x\right)=\frac{\left|S\right|}{\left|X\right|}$ - כלומר שכיחות x שווה לשכיחות הממוצעת במדגם - אזי $p_{MLE}\left(x\right)=\frac{1}{\left|X\right|}$
לכן: עבור $C\left(x\right)<\frac{\left|S\right|}{\left|X\right|}$ נקבל אומדן MLE קטן מאומדן התפלגות יוניפורמית:

$$C\left(x\right)<\frac{\left|S\right|}{\left|X\right|}\implies p_{MLE}\left(x\right)=\frac{C\left(x\right)}{\left|S\right|}<\frac{\frac{\left|S\right|}{\left|X\right|}}{\left|S\right|}=\frac{1}{\left|X\right|}$$

מסקנה - לxים כנ"ל יתקיים:

$$p_{Lid}\left(x\right)>_{PMLE}\left(x\right)$$

בניגוד להתנהגות מצופה משיטת discounting