

להחלקת לידסטון היה פרמטר אחד -  $\lambda$ . ככל שה  $\lambda$  יותר גדול ההחלקה יותר חזקה. לשיטות החלקה אחרות יכולים להיות עוד פרמטרים. איך מחליטים איזה פרמטרים לקבוע?

## כיול אמפירי של פרמטרים חיצוניים של שיטות החלקה/אומדן

### כיול על סמך מיקסום יעד

#### גישה א

נכיל את הפרמטר ע"י בדיקת טווח ערכים סביר ובדיקה איזה ערך מביא לביצועים האופטימליים באפליקציית היעד(למשל דיבור, תרגום וכו').  
זוהי גישה הנדסית פרופר - אין הרבה תיאוריה - אבל בפועל משתמשים בה הרבה.

#### גישה ב

יש מקרים שבהם נרצה קריטריון כללי לאיכות מודל ללא תלות באפליקציה מסוימת.  
אינטואיציה: יש לנו:

מדגם אימון: בניית המודל אומדן -  $p(x)$

מדגם מבחן:  $T = w_1, \dots, w_n$

נרצה אומדנים מדויקים ככל האפשר של  $p(x)$

אם  $p_1(x)$  יותר מדויק מ  $p_2(x)$ , אז מצפים ש  $p_1(T) > p_2(T)$

הרעיון הוא לנסות כמה הצבות פרמטרים, ולבדוק כל הצבה על מדגם מבחן.

שים  $\heartsuit$ : ההבדל בין הגישה הזו לגישה א' היא שבגישה א' בודקים את הביצועים של האפליקציה, ואילו כאן בודקים ישירות את המודל.

נעריך איכות של מודל שהפרמטרים שלו נאמדו על ידי סטטיסטיקות ממדגם אימון על ידי חישוב ההסתברות שהמודל נותן למדגם מבחן חדש גדול(מאותו מקור).

מתקיים: מודל שהאומדנים שלו "קרובים" יותר לערכים ההסתברותיים של הפרמטרים של המקור ייתן הסתברות גבוהה יותר למדגם המבחן.

#### לדוגמה

במודל מולטינומי(יוניגרם): מבחן:  $w_1, \dots, w_n$

$$p(w_1, \dots, w_n) = \prod_{i=1}^n p(w_i)$$

חישובית: • אם  $n$  גדול וננסה לחשב את  $\prod_{i=1}^n p(w_i)$  כנראה נקבל underflow. לכן נחשב לוג.

• במקום להסתכל על המכפלה נסתכל על השורש ה  $n$  שלה כדי לקבל ערך מנורמל שאינו תלוי ב  $n$ .  
- הממוצע הגיאומטרי של ההסתברויות במכפלה.

$$\log \sqrt[n]{p(w_1^n)} = \log \sqrt[n]{\prod_{i=1}^n p(w_i)} = \frac{1}{n} \log \prod_{i=1}^n p(w_i) = \frac{1}{n} \sum_{i=1}^n \log p(w_i)$$

פיתוח כנ"ל לכל מודל כפלי - למשל לטריגרם:

$$= \frac{1}{n} \sum_{i=1}^n \log p(w_i | w_{i-2}, w_{i-1})$$

## נבוכות (Perplexity)

כדי לקבל עבור כל מודל גודל אינטואיטיבי שמייצג את האיכות שלו, נחשב את perplexity (נבוכות):

$$\text{perplexity} = \frac{1}{\sqrt[n]{p(w_1^n)}}$$

אם ההתפלגות אחידה, אז  $p(w) = \frac{1}{|W|}$  ולכן  $\frac{1}{\sqrt[n]{p(w_1^n)}} = \frac{1}{\sqrt[n]{\left(\frac{1}{|W|}\right)^n}} = \frac{1}{\frac{1}{|W|}} = |W|$ .  
הפרפלקסיטי  $m$  של מודל מציין בחירה ששקולה ברמת אי הוודאות שלה לבחירת ערך ממודל יוניפורמי עם  $m$  ערכים שונים. ככל שהפרפלקסיטי נמוך יותר, המודל קרוב יותר להתפלגות האמיתית.

## חישוב הפרפלקסיטי

$$\frac{1}{\sqrt[n]{p(w_1^n)}} = 2^{\log \frac{1}{\sqrt[n]{p(w_1^n)}}}$$

למשל ביוניגרם:

$$= 2^{-\frac{1}{n} \sum_{i=1}^n \log p(w_i)}$$

## הערה

שיטות אלו טובות באופן כללי לאבלואציות שמשוות בין איכות של מודלים שונים (ולא רק לכיול פרמטרים).

## תכונות/בעיות בשיטת Lidstone

- קיימת תלות ב  $|X|$  (מספר הערכים האפשריים) - שדורשת לכייל את  $\lambda$  בהתאם. התלוד של האומד ב  $|X|$  לא משקפת את ההשפעה הנכונה שלו על האומדנים.
- רמת החלקה נקבעת לפי סכימה ליניארית, בעוד שנצפה להפחתה קטנה יותר יחסית למאורעות עם מונים גבוהים. (נצפה לסכימה לא ליניארית).

כללית לסכימת החלקה נדרשות שתי החלטות:

1. כמה מסת הסתברות להקצות למאורעות שלא נצפו
2. כמה הסתברות להפחית מכל מאורע שכן נצפה

נראה כעת שיטת החלקה אחרת, שמנסה לפתור את הבעיות האלה:

## שיטת החלקת Held-Out

גישה אמפירית להחלקה, ללא הנחות על התפלגות המקור

אינטואיציה נפריד את מדגם האימון ל- $S^T$  - המדגם שבו נשתמש לאימון - ו- $S^H$  - מדגם Held-Out.

- כדי להחליט כמה סיכוי להקצות למלים שלא היו במדגם האימון, נסתכל על אחוז המילים ב- $S^H$  שלא מופיעות ב- $S^T$  - זהו האומדן שלנו לסיכוי של מילים שלא הופיעו ב- $S^T$  להופיע בטקסט אמיתי, ונפרוס אותו על פני כל המילים של שלא ראינו במדגם.
- כדי להחליט כמה להפחית מהמילים שכן היו במדגם האימון, נסתכל על המלים שכן הופיעו ב- $S^T$  פעם אחת ונסתכל כמה פעמים הם הופיעו ב- $S^H$ , ולפי זה נחליט כמה להפחית מהם. אותו דבר עבור כל המלים שהופיעו פעמיים ב- $S^T$  וכן הלאה.

**הרעיון:** לקבוצת כל המאורעות שהופיעו  $r$  פעמים במדגם האימון נבדוק כמה פעמים בסה"כ חברי הקבוצה הופיעו ב- $S^H$ , ונחשב זאת כאומדן ממוצע לשכיחות המצופה שלהם במדגם אקראי חדש.

נסמן:	$N_r$	מספר המאורעות שהופיעו $r$ פעמים במדגם האימון
	$C^T()$	מונה ב- $S^T$
	$C^H()$	מונה ב- $S^H$
	$t_r$	מספר הפעמים הכולל שמאורעות בשכיחות $r$ ב- $S^T$ הופיעו ב- $S^H$ .

$$t_r = \sum_{x: C^T(x)=r} C^H(x)$$

$$P_{HO}(x|C^T(x)=r) = \frac{t_r}{|S^H| \cdot N_r}$$