

אומדן פרמטרים θ של מודל חבוי ללא נתונים מתוייגים (unsupervised)

מטרה: שערך של θ_{ML} , בהינתן מדגם של N תצפיות בלתי תלויות $y = y_1, \dots, y_N$:

$$\theta_{ML} = \operatorname{argmax}_{\theta} p(y; \theta) = \operatorname{argmax}_{\theta} \prod_{t=1}^N p(y_t; \theta)$$

- בעיות: יכולים להיות **המון** פרמטרים ב θ
- ברוב המודלים "המעשיים" - אין פתרון אנליטי, ולכן נדרש אלגוריתם חיפוש

קיים אלגוריתם איטרטיבי שנקרא EM (Expectation Maximization) שמחפש בכל איטרציה ערך חדש ל θ (אומדן θ), כך שמובטח שבכל איטרציה ערך הנראות $p(y; \hat{\theta})$ יעלה, עד התכנסות. לכן מובטחת התכנסות לנקודת מקסימום, ייתכן מקסימום לוקאלי.

אלגוריתם EM למודל עירוב היסטוגרמות

אתחול: ערך θ כלשהו (רנדומית, יוניפורמית, או "ניחוש מושכל"¹)

נחזור איטרטיבית:

1. שלב E : נחשב לכל מסמך y_t את $w_{ti} = p(x_i = x_t | y_t; \theta)$ - ההסתברות שהמסמך שייך לקטגוריה x_i . E=Expectation
2. שלב M : ניתן להסתכל על w_{ti} כתוחלת/הסתברות שאת הסטטיסטיקת y_t צריך לשייך ל x_i . M=Maximization

$$p(x_i) = \frac{\sum_{t=1}^N w_{ti}}{\sum_{j=1}^{|X|} \sum_{t=1}^N w_{tj}} = \frac{\sum_{t=1}^N w_{ti}}{N}$$

אינטואיטיבית: תוחלת מניית מספר מסמכי x_i מנורמל בסכום התוחלות לכל ערכי X ($N=|X|$)

$$p(w_k | x_i) = \frac{\sum_{t=1}^N n_{tk} \cdot w_{ti}}{\sum_{\ell=1}^V \sum_{t=1}^N n_{t\ell} \cdot w_{ti}} = \frac{\sum_{t=1}^N n_{tk} \cdot w_{ti}}{\sum_{t=1}^N w_{ti} \cdot n_t}$$

- כאשר: n_{tk} - שכיחות מילה w_k ב y_t
- n_t - אורך מסמך y_t

3. עצירה: כאשר $p(y; \theta)$ לא עולה.

לביצוע תנאי העצירה:

¹למשל - לאתחל את ההסתברות של מילה בכל הקטגוריות לפי ההסתברות שלה בלי חלוקה לקטגוריות

- נחשב לכל איטרציה את $p(y; \theta)$ (הערת debug: נבדוק שגדל מאיטרציה קודמת)
- נעצור אם שווה (פרקטית הבדל קטן) מאיטרציה קודמת (או איטרציות קודמות)

אי שוויון Jensen

זהו אי שוויון שמתקיים על פונקציות קמורות

הגדרה - פונקציה קמורה²

פונקציה $f(x)$ נקראת קמורה (Convex) אם לכל x_1, x_2 בתחום, ולכל $\alpha \in [0, 1]$ מתקיים:

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha \cdot f(x_1) + (1 - \alpha) \cdot f(x_2)$$

בכיוון ההפוך

פונקציה $f(x)$ נקראת קעורה (Concave) אם לכל x_1, x_2 בתחום, ולכל $\alpha \in [0, 1]$ מתקיים:

$$f(\alpha x_1 + (1 - \alpha)x_2) \geq \alpha \cdot f(x_1) + (1 - \alpha) \cdot f(x_2)$$

הרחבה לאוסף של נקודות

למספר כלשהו של נקודות x_1, \dots, x_n בתחום, ועבור $\alpha_1, \dots, \alpha_n \in [0, 1]$ כך $\sum_{i=1}^n \alpha_i = 1$, יתקיים בפונקציה קמורה:

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i \cdot f(x_i)$$

אינטרפרטציה הסתברותית

x_i ערכי משתנה מקרי דיסקרטי X

α_i הסתברות $p(x_i)$

נקבל עבור פונקציה קמורה:

$$f\left(\sum_{i=1}^n p(x_i) \cdot x_i\right) \leq \sum_{i=1}^n p(x_i) \cdot f(x_i)$$

$$f(E[x]) \leq E[f(x)]$$

במילים: אם f פונקציה קמורה, ונפעיל אותה על תוחלת, נקבל ערך יותר קטן מאשר אם נפעיל תוחלת על הפונקציה.

אנטרופיה יחסית (Relative Entropy)

נקרא גם Kullback-Leibler Divergence

מטרה: מדד לסטיה בין שתי התפלגויות

הגדרה

נתונות שתי התפלגויות דיסקרטיות p, q מעל n ערכים $p = p_1, \dots, p_n$ $q = q_1, \dots, q_n$. נגדיר:

$$D_{KL}(p||q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$$

משמעות בתורת האינפורמציה

כמות אובדן המידע (בביטים) כשמדלים את p ע"י q .

מקרי קצה

עבור $p = q$ - $D(p||q) = 0$
אם קיים i כך ש $q_i = 0$, אזי D_{KL} לא מוגדר (" ∞ ")

טענה

$$D(p||q) \geq 0$$

הוכחה

$$D(p||q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \leq \log \sum_i p_i \cdot \frac{q_i}{p_i} = \log_1 = 0$$

log is convex
Jensen's inequality



הצורה הכללית של אלגוריתם EM

המאמר המקורי משנת 1977:

Maximum Likelihood from Incomplete Data via the EM Algorithm
Dempster, Laird, Rubin

נלמד גרסא של הפיתוח:

Neal & Hinton, 98: A View of the EM Algorithm

המטרה: קירוב θ_{ML} , ע"י מציאת $\hat{\theta}$, עבור משתנים חבויים כשאין פתרון אנליטי, ע"י אלגוריתם איטרטיבי.

מניחים התפלגות $P(X, Y; \theta)$, חבוי X , גלוי Y ,
מחפשים:

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta} \log p(y; \theta) = \operatorname{argmax}_{\theta} \log \sum_{x \in X} p(x, y; \theta)$$

בדרך כלל אין פתרון אנליטי(ללוג של סכום) - נפתח שיטה איטרטיבית שבה בכל איטרציה ערך $p(y; \theta)$ יעלה.