מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 2015-12-27

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

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

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

$$\theta_{ML}=\underset{\theta}{\operatorname{argmax}}P\left(y;\theta\right)ssssss=\underset{\theta}{\operatorname{argmax}}\prod_{t=1}^{N}P\left(y_t;\theta\right)$$

בעיות:
- יכולים להיות **<u>המון</u>** פרמטרים בθ
- ברוב המודלים "המעשיים" - אין פתרון אנליטי, ולכן נדרש אלגוריתם חיפוש

קיים אלגוריתם איטרטיבי שנקרא EM(Expectation Maximization) שמחפש בכל איטרציה ערך חדש לθ(אומדן לθ), כך שמובטח שבכל איטרציה ערך הנראות $P\left(y;\hat{\theta}\right)$ 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 $w_{ti}=P\left(x_t=x_i|y_t;\theta\right)$ - ההסתברות שהמסמך שייך לקטגוריה x_i. E=Expectation
   ניתן להסתכל על w_{ti} כתוחלת/הסתברות שאת הסטטיסטיות מy_t צריך לשייך לx_i.
2. שלב הM: M=Maximization

$$P\left(x_i\right)=\frac{\sum_{t=1}^{N}w_{ti}}{\sum\limits_{j=1}^{|X|}\sum\limits_{t=1}^{N}w_{tj}}=\frac{\sum\limits_{t=1}^{N}w_{ti}}{N}$$

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

$$P\left(w_k|x_i\right)=\frac{\sum\limits_{t=1}^{N}n_{tk}\cdot w_{ti}}{\sum\limits_{\ell=1}^{V}\sum\limits_{t=1}^{N}n_{t\ell}\cdot w_{ti}}=\frac{\sum\limits_{t=1}^{N}n_{tk}\cdot w_{ti}}{\sum\limits_{t=1}^{N}w_{ti}\cdot n_t}$$

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

3. עצירה: כאשר $P\left(y;\theta\right)$ לא עולה.
   לביצוע תנאי העצירה:

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

- נחשב לכל איטרציה את $P\left(y;\theta\right)$
  (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 קמורה[2]

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

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

בכיוון ההפוך

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

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

הרחבה לאוסף של נקודות

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

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

אינטרפרטציה הסתברותית

x_i	ערכי משתנה מקרי דיסקרטי X
α_i	הסתברות $p\left(x_i\right)$

נקבל עבור פונקציה קמורה:

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

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

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

[2] כשבודקים קעירות/קמירות מסתכלים על הפונקציה מלמטה

אנטרופיה יחסית (Relative Entropy)

נקרא גם Kullback-Leibler Divergence

מטרה: מדד לסטיה בין שתי התפלגויות

הגדרה

נתונות שתי התפלגויות דיסקרטיות q,p מעל n ערכים $\begin{array}{c} p=p_1,\dots p_n \\ q=q_1,\dots q_n \end{array}$. נגדיר:

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

משמעות בתורת האינפורמציה

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

מקרי קצה

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

טענה

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

הוכחה

$$D\left(p\|q\right)=\sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}\underset{\substack{\text{log is convex}\\ \text{Jensen's inequality}}}{\leq}\log\sum_{i}\cancel{p_i}\cdot\frac{q_i}{\cancel{p_i}}=\log_1=0$$

■

הצורה הכללית של אלגוריתם EM

המאמר המקורי משנת 1977:

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

נלמד גירסא 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 $P(X,Y;\theta)$, X-חבוי, Y-גלוי
מחפשים:

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

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