מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 20-12-2015

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

עירוב היסטוגרמות

סימונים

x_i	קטגוריות (למשל בדסק חדשות - מסמך ספורט, מסמך אקטואליה וכו')
y_t	מסמך - $y_1, y_2, \ldots$
V	גודל המילון
w_k	המילה הk במילון
n_{tk}	שכיחות מילה w_k בy_t
θ	אוסף הפרמטרים
$P(x_i)$	הסתברות שמסמך יהיה עם ערך $X = x_i$

השאלות שנתעניין

1. התסברות תצפית מסויימת y_t:

$$P(y_t;\theta) = \sum_{i=1}^{|X|} P(x_i) \cdot P(y_t|x_i)$$

בעירוב היסטוגרמות זה שווה ל:

$$= \sum_{i=1}^{|X|} P(x_i) \cdot \prod_{k=1}^{V} P(w_k|x_i)^{n_{tk}}$$

2. "סיווג" - הסתברות לכל ערך חבוי אפשרי:

$$P(X = x_i|y_t;\theta) = \frac{P(y_t|x_i)\,P(x_i)}{P(y_t)}$$

בעירוב היסטוגרמות זה שווה ל:

$$= \frac{P(x_i) \cdot \prod_{k=1}^{V} P(w_k|x_i)^{n_{pk}}}{\sum_{j=1}^{|X|} P(x_j) \prod_{k=1}^{V} P(w_k|x_j)^{n_{tk}}}$$

במקרה שרוצים להחליט על סיווג לערך הסביר ביותר i^*:

$$i^* = \underset{i}{\operatorname{argmax}} P(y_t|x_i)\,P(x_i) = \underset{i}{\operatorname{argmax}} \left(\log P(x_i) + \sum_{k=1}^{V} n_{tk} \cdot \log P(w_k|x_i) \right)$$

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 20-12-2015

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

אומדן θ במקרה המבוקר Supervised

נתון מדגם של מסמכים y_t שלכל אחד מהם תוייג ערך x_i המתאים(ערך הקטגוריה). לדוגמה, אומדן עם החלקת Lidstone:

$$P(x_i) = \frac{\lambda + \text{number of } x_i \text{ documents}}{\lambda \cdot |X| + \text{total number of documents}}$$

הערה: במכנה מכפילים את λ ב$|X|$ כי יש $|X|$ קטגוריות

$$P(w_k|x_i) = \frac{\lambda + \text{number of occurences of } w_k \text{ in } x_i \text{ documents}}{\lambda \cdot V + \text{total number of } x_i \text{ documents}}$$

הערה: במכנה מכפילים את λ בV כי w_k נבחר מתוך V מילים

מודלים אחרים

את המודלים האלו הראנו עם מודל מסמך מולטינומי - אבל אפשר לשתול גם מודלים אחרים בתוך המודל הזה. הפיתוח של הנוסחאות יראה כמובן אחרת.

מודל חבוי עם מודל מסמך של Multiple Bernoulli

במודל זה: מסמך מיוצג כקבוצת המילים(ללא התייחסות לשכיחות). ניתן לייצג כוקטור בינארי באורך V) במקרה זה:

$$P(y_t|x_i;\theta) = \prod_{w \in y_t} P(w|x_i) \cdot \prod_{w \notin y_t} (1 - p(w|x_i))$$

במודל חבוי - נשתמש בנוסחה זו.

אומדן הפרמטרים במקרה הSupervised

אם קודם היה מודל ברנולי אחד לכל המסמכים - עכשיו יש לנו מודל ברנולי נפרד לכל קטגוריה!

$P(x_i)$ כמו קודם

$$P(w_k|x_i) = \frac{\lambda + \text{number of } x_i \text{ documents where } w_k \text{ appears}}{\lambda \cdot 2 + \text{number of } x_i \text{ documents}}$$

הערה: במכנה מכפילים את λ ב2 כי לכל מילה w_k יש 2 אפשרויות - או שהיא מופיעה או שהיא לא מופיעה

מודל חבוי במקרה הלא מבוקר Unsupervised

במקרה הזה, לא נתון מדגם מתוייג, אבל עדיין מניחים שהתצפיות נוצרות מהתפלגויות שונות, כתלות בערך חבוי מX.

מטרות לשימוש במודל חבוי Unsupervised:

1. כשיש מטרה יישומית לאשכול(clustering) של התצפיות, נרצה לשייך כל y_t לx_i הכי סביר.

   למשל: להצגת מסמכים לפי קבוצות

2. מידול יותר מדוייק של יצירת הנתונים, שייתן נראות גבוהה יותר לתצפיות. נחפש θ_{ML}

מקליד: עידן אריה
מרצה: פרופ' עידו דגן
תאריך: 20-12-2015

מודלים הסתברותיים יישומיים במדעי המחשב
89-919-01

דוגמה לכך שמודל חבוי מאפשר אומדן נראות גבוה יותר

$w^1, w^2, \ldots w^{10} \quad V = 10$

$x_1, \ldots x_{10} \quad |X| = 10$

תצפית: 10 מסמכים - $y_1, \ldots y_{10}$

בכל d_i - מופיעה רק המילה w^i 10 פעמים

מודל ללא משתנה חבוי: MLE

יש לנו סה"כ 100 מילים(10 מסמכים * 10 מילים בכל מסמך), וכל מילה מופיעה סה"כ 10 פעמים(מסמך אחד * 10 פעמים), לכן האומדן לכל מילה הוא:

$$\forall_i P\left(w^i\right) = \frac{10}{100} = \frac{1}{10}$$

לפי זה, ההסתברות לקבל מסמך ספציפי לפי ההסתברויות הנ"ל:

$$\forall_{i=1,\ldots 10} P\left(y_i\right) = P\left(w^i\right)^{10} = 10^{-10}$$

מודל חבוי: MLE

נגדיר 10 התפלגויות, כך שכל אחת תתאים למסמך אחר:

$$\forall_i P\left(w^i \middle| x_i\right) = 1 \qquad \forall_{j \neq i} P\left(w^j \middle| x_i\right) = 0$$

$$\forall_i P\left(x_i\right) = \frac{1}{10}$$

$$\forall_i P\left(y_i | x_i\right) = 1 \qquad \forall_{j \neq i} P\left(y_i | x_j\right)$$

$$\forall_i P\left(y_i\right) = \sum_j P\left(x_j\right) \cdot P\left(y_i | x_j\right) = \underbrace{\frac{1}{10} \cdot 1}_{i=j} + \underbrace{\left(\frac{1}{10} \cdot 0\right) \cdot 9}_{i \neq j} = \frac{1}{10}$$

נשים לב

הצלחנו לייצר נראות יותר גבוהה - אבל המחיר הוא שהשתמשנו ביותר פרמטרים. בהתפלגות ללא המשתנה החבוי היו לנו 10 פרמטרים(הסתברות לכל מילה), אבל כאשר הכנסנו את המשתנה החבוי היינו צריכים 100 פרמטרים(הסתברות לכל מילה בכל מסמך).

הוספת כמות גדולה יותר של פרמטרים במודל החבוי אפשרה למדל בצורה הדוקה יותר את התצפית ולקבל נראות גבוהה יותר.

צריך לשים לב שלא לוקחים יותר מדי פרמטרים, כי אז עלולים לעשות overfitting וליצור מודל שמתאים בדיוק למדגם האימון.