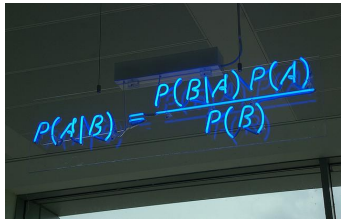
מבוא לאיחזור מידע



מעבדה 4 - סיווג

מטלות המעבדה - סיווג בעזרת מסווג בייסיאני

נתונים למעבדה זו : from sklearn.datasets import fetch_20newsgroups

ניתן לבחור מספר קטגוריות בודד לסיווג.

1. יש לממש מסווג בייסיאני בעזרת ספריות סטנדרטיות, numpy, pandas .
   a. מולטי נומינלי
   b. מולטי ווריאנטי
2. יש לטפל במסמכים בהיבט העיבוד המקדים (כמו ייצוג של BOW - count vectors) שימו לב ניתן **ורצוי** להיעזר בכלים שפותחו בעבר (ניתן גם להיעזר בספריות פייתון המיועדות לכך from sklearn.feature_extraction.text import CountVectorizer)
3. יש לסווג את המסמכים בעזרת כל מסווג (לא לשכוח חלוקה ל train ו test).
4. יש להציג מדדי הערכה לכל מסווג (על פי הנלמד בהרצאה) ניתן להשתמש בספריות פייתון המיועדות לכך (כמו sklearn.metrics).
5. חיזרו על סעיפים 2-4 בעזרת מסווגים מוכנים מאחת מספריות פייתון לעיבוד שפה

סיווג בייסיאני מולטינומינלי מול מולטי ווריאנטי

סיווג זה יהיה רלוונטי כאשר יש פחות משמעות לתדירות המילים, אלא רק הנוכחות שלהם במסמך.
לדוגמא זיהוי רגש במסמך.
בחישוב בדיקת קיום או אי קיום של מושג במסמך ביחס לסיווג (בדומה לתדירות מסמכים).
בדוגמא סין מופיעה בשלושה מסמכי C.

Example (multivariate)

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$$
$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0+1)/(3+2) = 1/5$$
$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1+1)/(3+2) = 2/5$$
$$\hat{P}(\text{Chinese}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) = (0+1)/(1+2) = 1/3$$

$$\hat{P}(c) = 3/4$$
$$\hat{P}(\overline{c}) = 1/4$$

$$\begin{aligned}\hat{P}(c|d_5) &\propto \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\\ &\quad\cdot(1-\hat{P}(\text{Beijing}|c))\cdot(1-\hat{P}(\text{Shanghai}|c))\cdot(1-\hat{P}(\text{Macao}|c))\\ &= 3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\\ &\approx 0.005\end{aligned}$$

$$\boxed{\begin{aligned}\hat{P}(\overline{c}|d_5) &\propto 1/4\cdot 2/3\cdot 2/3\cdot 2/3\cdot(1-1/3)\cdot(1-1/3)\cdot(1-1/3)\\ &\approx 0.022\end{aligned}}$$

סיווג בייסיאני מולטינומינלי מול מולטי ווריאנטי

סיווג זה יהיה רלוונטי כאשר רוצים לתת משקל לתדירות המילים במסמך.
ספירת כמות המופעים של המושג בכל מסמך ביחס לסיווג (סכום של TF).
בדוגמא סין מופיעה פעמיים במסמך 1, פעמיים במסמך 2 ופעם נוספת בשלישי - סהכ חמש פעמים.

Example (multinomial)

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(\text{Chinese}|c) = (5+1)/(8+6) = 6/14 = 3/7$$
$$\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) = (0+1)/(8+6) = 1/14$$
$$\hat{P}(\text{Chinese}|\bar{c}) = (1+1)/(3+6) = 2/9$$
$$\hat{P}(\text{Tokyo}|\bar{c}) = \hat{P}(\text{Japan}|\bar{c}) = (1+1)/(3+6) = 2/9$$

$$\hat{P}(c) = 3/4$$
$$\hat{P}(\bar{c}) = 1/4$$

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003.$$
$$\hat{P}(\bar{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001.$$