

# מבוא לאיחזור מידע

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

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. אנחנו נשתמש בסוגים שונים של NB לצורך בניית המודלים.

2. יש לבחור מספריות מוכנות לעיבוד שפה

a. מולטי נומינלי

b. מולטי ווריאנטי

3. 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`

4. יש לסווג את המסמכים בעזרת כל מסווג (לא לשכוח חלוקה ל `test` ו `train` ).

5. יש להציג מדדי הערכה לכל מסווג (על פי הנלמד בהרצאה) ניתן להשתמש בספריות פייתון

המיועדות לכך (כמו `sklearn.metrics` ).

# סיווג בייסיאני מולטינומינלי מול מולטי ווריאנטי

## 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	?

$$\hat{P}(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$$

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

$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0+1)/(3+2) = 1/5$$

$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1+1)/(3+2) = 2/5$$

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

$$\hat{P}(\text{Chinese}|\bar{c}) = (1+1)/(1+2) = 2/3$$

$$\hat{P}(\text{Japan}|\bar{c}) = \hat{P}(\text{Tokyo}|\bar{c}) = (1+1)/(1+2) = 2/3$$

$$\hat{P}(\text{Beijing}|\bar{c}) = \hat{P}(\text{Macao}|\bar{c}) = \hat{P}(\text{Shanghai}|\bar{c}) = (0+1)/(1+2) = 1/3$$

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

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

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

# סיווג בייסיאני מולטינומינלי מול מולטי ווריאנטי

## Introduction to Information Retrieval

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	?

$$\begin{aligned}\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 & \hat{P}(c) &= 3/4 \\ \hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 & & \\ \hat{P}(\text{Chinese}|\bar{c}) &= (1+1)/(3+6) = 2/9 & \hat{P}(\bar{c}) &= 1/4 \\ \hat{P}(\text{Tokyo}|\bar{c}) = \hat{P}(\text{Japan}|\bar{c}) &= (1+1)/(3+6) = 2/9 & & \end{aligned}$$

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

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