۴ـ یک سند مانند d ممکنه چندین کلاس داشته باشد. مثلاً هم هنری باشد و هم اقتصادی. با فرض اینکه سند d را داریم، می‌خواهیم کلاسی که برای آن سند، احتمال بیشتری دارد را به عنوان کلاس آن سند در نظر بگیریم. طبق قانون Bayes داریم:

$$p(c|d) = \frac{p(d|c)\,p(c)}{p(d)}$$

چون هدف یافتن کلاسی با بیشترین احتمال است و از طرفی، سند ثابت است، در محاسبات مخرج کسر را محاسبه نمی‌کنیم

$$c_{MAP} = \arg\max p(d|c) \times p(c)$$

سند d، در اینجا از فیچرهای کلمه تشکیل شده است. ←

$$c_{MAP} = \arg\max p(x_1,\dots,x_n|c) \times p(c)$$

پس اگر فرض کنیم تعداد کلمات n تا است: $x_1,\dots,x_n$، برای ترتیب این کلمات $|x|^n$ حالت مختلف داریم و چون c تا کلاس داریم تعداد پارامترها $O(|x|^n \cdot |c|)$ خواهد بود و این یک order نمایی است. با در نظر گرفتن شرط مستقل بودن کلمات از یکدیگر (که در عمل همیشه اینطور نیست و این یک فرض غیرواقعی است)، order را به طور قابل ملاحظه‌ای کاهش می‌دهیم.

(فرض دیگه‌ای هم داریم اونم اینکه کلمات به صورت bag of word هستند یعنی فرض می‌کنیم ترتیب کلمات مهم نیست).

$$c_{MAP} = \arg\max p(x_1,\dots,x_n|c) \times p(c)$$

با فرض استقلال کلمات

$$c_{NB} = \arg\max p(c_j) \prod_{x \in X} p(x|c)$$

$p(x_1,\dots,x_n|c)$ استقلال کلمات → $p(x_1|c) \times p(x_2|c) \dots \times p(x_n|c)$

اگه فرض استقلال کلمات در قانون Bayes را در نظر بگیریم بهش می‌گن Naive Bayes

پس علت تفاوت این ۲ فرمول اینست که فرمول دوم با فرض مستقل بودن کلمات است تا تعداد پارامترها و حجم مورد نیاز به شدت کاهش یابد.

۵- از کل مجموعه‌ی training set، vocabulary را استخراج می‌کنیم، به ازای هر کدام از کلاس‌ها، همه‌ی داکیومنت‌های متعلق به آن کلاس را می‌شماریم و تقسیم بر تعداد کل document ها می‌کنیم.

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

احتمال هر کلمه در کلاس c_j را به دست می‌آوریم، یعنی تعداد آن کلمه در کلاس c_j، تقسیم بر تعداد همه‌ی کلمات کلاس c_j:

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

با توجه به فرمول بالا، اگر یکی از کلمات test set در داکیومنت‌های train مربوط به آن موضوع دیده نشده، احتمال تعلق سند به آن موضوع ۰ می‌شود. حال باید به گونه‌ای باشد که کلمات دیده نشده را نیز احتمال وقوع دهد. یعنی ممکن است کلمه‌ای در train نباشد ولی دلیل نمی‌شود که به آن موضوع تعلق نداشته باشد. برای جلوگیری از ۰ شدن احتمالات، از روش لاپلاس در رده بندی Bayes استفاده می‌کنیم:

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j) + 1}{\left[\sum_{w \in V} count(w, c_j)\right] + |V|} = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} \left(count(w, c_j) + 1\right)}$$

۶-

$$\hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$$

$P(c) = \frac{3}{4}$ ← احتمال کلاس c (label c)

$P(j) = \frac{1}{4}$ ← احتمال کلاس j (label j)

$|V| = 6$ ← تعداد کل کلمات که متفاوت هستند

chinese chinese chinese Tokyo japan

$P(chinese|c) = \frac{5+1}{8+6} = \frac{3}{7}$

$P(chinese|j) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(Tokyo|c) = \frac{0+1}{8+6} = \frac{1}{14}$

$P(Tokyo|j) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(japan|c) = \frac{0+1}{8+6} = \frac{1}{14}$

$P(japan|j) = \frac{1+1}{3+6} = \frac{2}{9}$

$P(c|d) =$

$\frac{3}{4} \times \frac{1}{14} \times \frac{1}{14} \times (\frac{3}{7})^3$ ← ۳ بار واژه این کلمه

$\frac{3}{4}$ ← احتمال خود کلاس c

$\frac{1}{14}$ ← $P(japan|c)$

$\frac{1}{14}$ ← $P(Tokyo|c)$

$(\frac{3}{7})^3$ ← $P(chinese|c)$

$\simeq 0.0003$ ✓

کلاس c انتخاب می‌شود

$P(j|d) = \frac{1}{4} \times \frac{2}{9} \times \frac{2}{9} \times (\frac{2}{9})^3 \simeq 0.0001$

۷-

$P(\text{سیاسی بودن} \mid \text{پارلمان و مجلس و پرسپولیس}) = ؟$

ضرب صورت‌ها = ۹

$P(\text{پارلمان و مجلس و پرسپولیس و سیاسی}) = \frac{1}{8} \times \frac{3}{8} \times \frac{1}{8} \times \frac{3}{7}$

ضرب صورت‌ها = ۰

$P(\text{پارلمان و مجلس و پرسپولیس و ورزشی}) = \frac{3}{8} \times \frac{0}{8} \times \frac{3}{8} \times \frac{4}{7}$

$P(\text{سیاسی بودن} \mid \text{پارلمان و مجلس و پرسپولیس}) = \frac{9}{9+0} = \frac{9}{9} = 1$

$P(\text{ورزشی} \mid \text{پارلمان و مجلس و پرسپولیس}) = \frac{0}{9+0} = \frac{0}{9} = 0$