Document Classification Problem

We have k classes of documents.

(Example: class 1 : spam, class 2 : important, class 3: personal)
class 4 : work.

$k=4$

Dataset: class 1 has 5000 documents.
has class 2 " 6000 "
25K " 3 " 4000 "
documents " 4 " 10000 " .

Prior probabilities: $\pi_1 = \dfrac{5000}{20000} = \dfrac{5}{25}, \cdots \pi_4 = \dfrac{10}{25}$

$\pi_2 = \dfrac{6}{25}, \quad \pi_3 = \dfrac{4}{25}.$

$V = \{\text{word1}, \text{word2}, \ldots, \text{word}_{|V|}\}$ has $|V|$ elements.

$\bar{p}^{(1)} = (P_{11}, \cdots P_{1|V|})$ is the model for class 1.

$\bar{p}^{(2)} = (P_{21}, \cdots, P_{2|V|})$ " " " " 2.

$$\hat{P}_{21} = \frac{\#\text{ of word 1 in class 2}}{(\#\text{ of word1} + \cdots + \#\text{ word}|V|)\text{ in class 2}}$$

$\bar{p}^{(3)} = (P_{31}, \cdots, P_{3|V|})$, $\quad \sum_{i=1}^{|V|} P_{3i} = 1$

$\bar{p}^{(4)} = (P_{41}, \cdots, P_{4|V|})$, $\quad \sum_{i=1}^{|V|} P_{4i} = 1$

$$\hat{P}_{43} = \frac{\#\text{ of word3 in class 4}}{(\#\text{ of word1} + \cdots + \#\text{ word}|V|)\text{ in class 4}}$$

We have 4 models:

Model 1 for class 1

$$\text{Likelihood}_1 = p_{11}^{x_1} \cdots \cdots p_{1|V|}^{x_{|V|}}$$

Model 2 for class 2

$$\text{Likelihood}_2(x_1, x_2, \cdots, x_{|V|}) = p_{21}^{x_1} p_{22}^{x_2} \cdots p_{2|V|}^{x_{|V|}}$$

Model 3 〃 $_3($ 〃 $) = p_{31}^{x_1} \cdots\cdots p_{3|V|}^{x_{|V|}}$

Model 4 〃 $_4($ 〃 $) = p_{41}^{x_1} \cdots\cdots p_{4|V|}^{x_{|V|}}$

Training process is over

<u>Testing</u> (Classification or Recognition) phase

Given a document

word1

word1, word2
word1, word2
word3, word4
word5, word6

$$\xrightarrow[\text{extraction}]{\text{feature}} \quad x = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$|V| = 6$

Feature vector is obtained using the "bag of words" model.

```
for $i = 1:4$
    $\alpha_i = \pi_i \, p_{i1}^{3} p_{i2}^{2} p_{i3}^{1} p_{i4}^{1} p_{i5}^{1} p_{i6}^{1}$
end.
```

Answer = pick the i that gives the max probability (likelihood).

$$\boxed{\text{Class Answer} = \arg\max_i (\alpha_i)}$$