Paper Code	Examiner	Department	Office
INT309	Fangyu Wu	Department of Intelligence Science	SD555

Based on the data in Table 1 below:

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

1. Estimate a multinomial Naive Bayes classifier, and apply the classifier to the test document.

SOLUTION. (i) $\hat{P}(c) = \hat{P}(\overline{c}) = 1/2$. The vocabulary has 7 terms: Japan, Macao, Osaka, Sapporo, Shanghai, Taipei, Taiwan. There are 5 tokens in the concatenation of all c documents. There are 5 tokens in the concatenation of all $\overline{c}$ documents. Thus, the denominators have the form (5+7). $\hat{P}(\text{Taiwan}|c) = (2+1)/(5+7) = 1/4$, $\hat{P}(\text{Taiwan}|\overline{c}) = (1+1)/(5+7) = 1/6$, $\hat{P}(\text{Sapporo}|c) = (0+1)/(5+7) = 1/12$, $\hat{P}(\text{Sapporo}|\overline{c}) = (2+1)/(5+7) = 1/4$,
(ii) We then get $\hat{P}(c|d) \propto 1/2 \cdot (1/4)^2 \cdot 1/12 = 1/(2^7 \cdot 3) \approx 0.00260$ and $\hat{P}(\overline{c}|d) \propto 1/2 \cdot (1/6)^2 \cdot (1/4) = 1/(2^5 \cdot 3^2) \approx 0.00347$. $\hat{P}(c|d)/\hat{P}(\overline{c}|d) = 3/4$. Thus, the classifier assigns the test document to $\overline{c}$ = *not-China*.

2. Estimate a Bernoulli Naive Bayes classifier and apply the classifier to the test document.

c. Estimating parameters of a binomial Naive Bayes classifier:

$p(c=\text{China})=0.5$, $p(c=\text{not China})=0.5$

$p(\text{Taiwan}|c=\text{China})=\frac{2+1}{2+2}=\frac{3}{4}$,

$p(\text{Sapparo}|c=\text{China})=p(\text{Japan}|c=\text{China})=p(\text{Osaka}|c=\text{China})=\frac{0+1}{2+2}=\frac{1}{4}$

$p(\text{Macao}|c=\text{China})=p(\text{Shanghai}|c=\text{China})=p(\text{Taipei}|c=\text{China})=\frac{1+1}{2+2}=\frac{1}{2}$

$p(\text{Sapparo}|c=\text{not China})=\frac{3}{4}$,

$p(\text{Taiwan}|c=\text{not China})=p(\text{Japan}|c=\text{not China})=p(\text{Osaka}|c=\text{not China})=\frac{1+1}{2+2}=\frac{1}{2}$

$p(\text{Taipei}|c=\text{not China})=p(\text{Shanghai}|c=\text{not China})=p(\text{Macao}|c=\text{not China})=\frac{1}{4}$

d. $p(\text{test set}|c=\text{China})=\frac{3}{4}*\frac{1}{4}*\left(\frac{3}{4}\right)^2*\left(\frac{1}{2}\right)^3=\frac{27}{2^{11}}$,

$p(\text{test set}|c=\text{not China})=\frac{1}{2}*\frac{3}{4}*\left(\frac{1}{2}\right)^2*\left(\frac{3}{4}\right)^3=\frac{81}{2^{11}}$

$\Rightarrow$ test set belongs to class "not china"

3. Compute the tf-idf vector (normalized) representations of the documents in the table below. By using Rocchio Classication Algorithm to determine the class of the document.

Solution: We measure the distance between the tf-idf query vector (normalized) and each centroid, and choose the class with the smallest distance. With normalization, the query is classified as not belonging to class c. If we only use raw tf query vector (normalized), no idf, the query is classified as belonging to class c.

——— *End of paper* ———