I R Final 610721204 陳文威

1. Naive Bayes text classification

(i) Bernoulli model

Document Term Matrix : (Binary)

	Hualien	Taiwan	Sapporo	Shanghai	Japan	class
d_1	1	1	0	0	0	Y
d_2	1	1	0	0	0	Y
d_3	0	0	1	1	0	Y
d_4	1	0	1	0	1	N
d_5	0	1	1	0	1	N
d_6	1	1	0	0	0	?

class : YES $\Rightarrow \frac{3}{5} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} \times \frac{3+1}{3+2} = 0.062\ldots$

class : No $\Rightarrow \frac{2}{5} \times \frac{1+1}{2+2} \times \frac{1+1}{2+2} \times \frac{0+1}{2+2} \times \frac{2+1}{2+2} \times \frac{0+1}{2+2} = 0.004\ldots$

Ans : class label : YES

(ii) Multinomial model

Label Term Matrix : (count)

	Hualien	Taiwan	Sapporo	Shanghai	Japan
YES	3	3	1	1	0
NO	1	1	2	0	3

class : YES $\Rightarrow \frac{3}{5} \times \frac{3+1}{8+5} \times \frac{3+1}{8+5} \times \frac{3+1}{8+5} = 0.017\ldots$

class No $\Rightarrow \frac{2}{5} \times \frac{1+1}{7+5} \times \frac{1+1}{8+5} \times \frac{1+1}{8+5} = 0.001\ldots$

ANS : class label : YES

2.

(i)

M_1 =

	①	②	③	④
①	0	$\frac{1}{2}$	$\frac{1}{2}$	0
②	$\frac{1}{2}$	0	0	$\frac{1}{2}$
③	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
④	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0

M_2 :

	①	②	③	④
①	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
②	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
③	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
④	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$0.8\,M_1 + 0.2\,M_2$ =

	①	②	③	④
①	$\frac{1}{20}$	$\frac{9}{20}$	$\frac{9}{20}$	$\frac{1}{20}$
②	$\frac{9}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{9}{20}$
③	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
④	$\frac{19}{60}$	$\frac{19}{60}$	$\frac{19}{60}$	$\frac{1}{20}$

#

(ii)

step 0 : $\langle 1, 0, 0, 0 \rangle$

step 1 : $\langle \frac{1}{20}, \frac{9}{20}, \frac{9}{20}, \frac{1}{20} \rangle$

step 2 : $\langle \frac{1}{20}\times\frac{1}{20} + \frac{9}{20}\times\frac{9}{20} + \frac{9}{20}\times\frac{1}{4} + \frac{1}{20}\times\frac{19}{60}$,

$\frac{1}{20}\times\frac{9}{20} + \frac{9}{20}\times\frac{1}{20} + \frac{9}{20}\times\frac{1}{4} + \frac{1}{20}\times\frac{19}{60}$,

$\frac{1}{20}\times\frac{9}{20} + \frac{9}{20}\times\frac{1}{20} + \frac{9}{20}\times\frac{1}{4} + \frac{1}{20}\times\frac{19}{60}$,

$\frac{1}{20}\times\frac{1}{20} + \frac{9}{20}\times\frac{9}{20} + \frac{9}{20}\times\frac{1}{4} + \frac{1}{20}\times\frac{1}{20} \rangle$

(iii)

$\langle 1, 0, 0, 0 \rangle \times M^{10}$

3. $n = 20$

Total number of pairs $= 20 \times 19 / 2 = 190$

$TP + FP = C_2^7 + C_2^8 + C_2^5 = 59$

$TP = C_2^4 + C_2^3 + C_2^5 + C_2^2 + C_2^3 + C_2^2 = 24$

$FP = 59 - 24 = 35$

$FN + TN = 190 - 59 = 131$

$FN = (4 \times 3 + 1 \times 2) + (3 \times 5) + (2 \times 3) = 35$

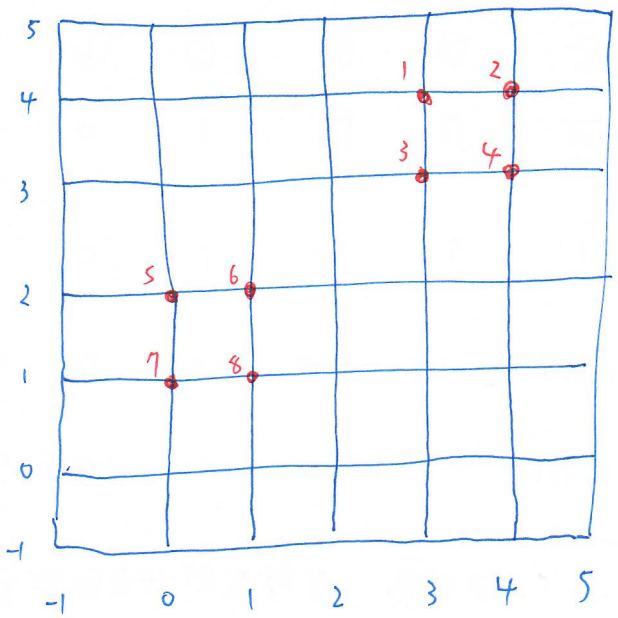
$TN = 131 - 35 = 96$

Confusion Matrix :

24 (TP)	35 (FN)
35 (FP)	96 (TN)

Rand Index : $\frac{24+96}{24+35+35+96} = \frac{120}{190} = 0.63$ #

4. (i) My idea : 使用 PCA 降維，並人工選取起始點。



example : 當分 2 群時，可以選 2，7 當作 seed。

(ii) K - means ++

step 1 : Randomly select the first centroid from the data points.

step 2 : For each data point compute it's distance from the nearest, previously choosen centroid.

step 3 : Select the next centroid from the data points such that the probability of chosing a point as centroid is directly proportional to it's distance from the nearest, previously chosen centroid.

step 4 : Repeat steps 2 and step 3 untill K centroids have been sampled.

example：以 step 1 選 6 號為例：

	①	②	③	④	⑤	⑥	⑦	⑧
$D(x)$	$2\sqrt{2}$	$\sqrt{13}$	$\sqrt{5}$	$\sqrt{10}$	1	0	$\sqrt{2}$	1
$D(x)^2$	8	13	5	10	1	0	2	1
$p(x)$	0.2	0.325	0.125	0.25	0.025	0	0.05	0.025
sum	0.2	0.525	0.65	0.9	0.925	0.925	0.975	1

隨機產生一個 0-1 之間的數，選擇區間代表的點當作 seed，

ex：①的區間為 0 - 0.2，②的區間為 0.2 - 0.525。

而①②③④就佔了整體的 90%，確實與現有的⑥較遠。