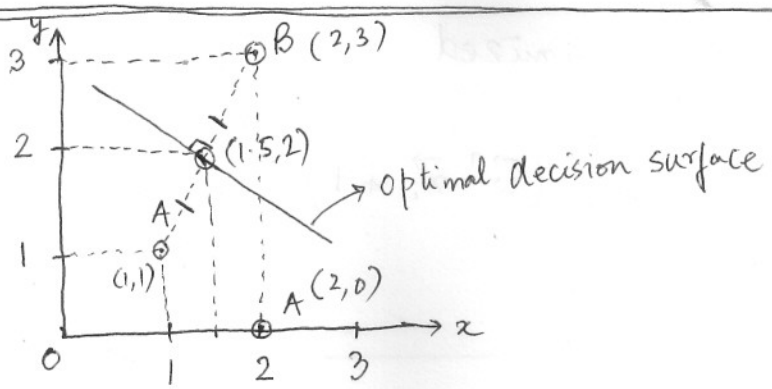
INFORMATION RETRIEVAL (CS60092)

Computer Science and Engineering, Indian Institute of Technology Kharagpur

End - Semester Examination

1.



1.(a) With the standard constraint that sign $(y_i (\vec{w}^T \vec{x}_i + b)) \geq 1$, we seek to minimize $|\vec{w}|$.

This happens when this constraint is satisfied with equality by the 2 support vectors. Further we know that the solution is $\vec{w} = (a, 2a)$ for some a. So we have that

$a + 2a + b = -1$

$2a + 6a + b = 1$

$\therefore$ $a = 2/5$ and $b = -11/5$. So the optimal hyperplane is given by $\boxed{\vec{w} = (2/5, 4/5) \text{ and } b = -11/5}$. Ans.

1.(b) Margin $\rho = \dfrac{2}{|\vec{w}|} = \dfrac{2}{\sqrt{\frac{4}{25} + \frac{16}{25}}} = \dfrac{2}{\frac{2\sqrt{5}}{5}} = \boxed{\sqrt{5}}$ Ans.

1.(c) See figure above.

---

1. (d) The required ~~formalized~~ formulation of the SVM optimization problem with slack variables is:

Find $\vec{w}$, b, and $\xi_i \geq 0$ such that:

- $\frac{1}{2} \vec{w}^T \vec{w} + C \sum_i \xi_i$ is minimized
- and for all $\{(\vec{x}_i, y_i)\}$, $y_i (\vec{w}^T \vec{x}_i + b) \geq 1 - \xi_i$

---

1. (e) The required dual problem for soft margin classification becomes:

Find $\alpha_1, \ldots, \alpha_N$ such that $\sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j$ is maximized, and

- $\sum_i \alpha_i y_i = 0$
- $0 \leq \alpha_i \leq C$ for all $1 \leq i \leq N$

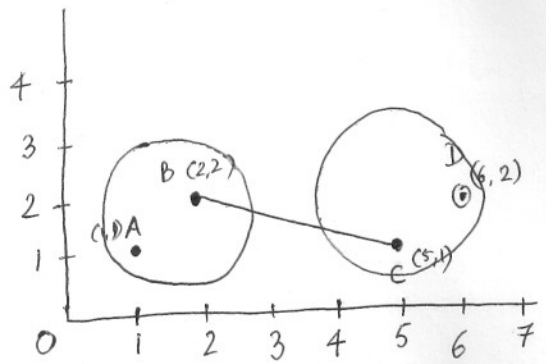
---

1. (f) Kernel functions.

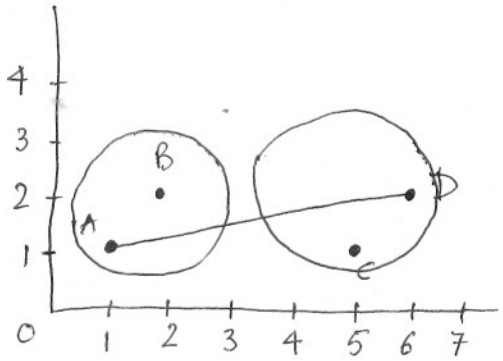
---

2. (a) A (1, 1), B (2, 2), C (5, 1), D (6, 2).

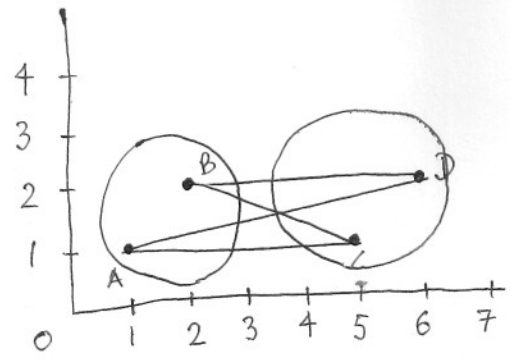
Cluster 1 Cluster 2



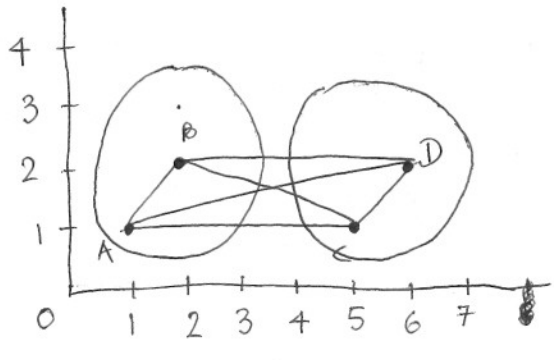
<i>



<ii>



<iii>



<iv>

<i> Similarity $= \sqrt{(2-5)^2 + (2-1)^2} = \sqrt{9+1} = \boxed{3.162}$ Ans.

<ii> Similarity $= \sqrt{(1-6)^2 + (1-2)^2} = \sqrt{25+1} = \boxed{5.099}$ Ans.

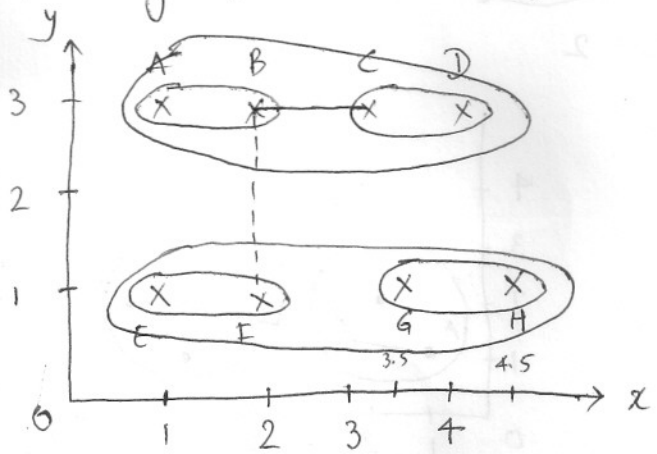
<iii> Similarity $= \dfrac{|AC| + |AD| + |BC| + |BD|}{4} = \dfrac{4 + 5.099 + 3.162 + 4}{4}$

$= \boxed{4.065}$ Ans.

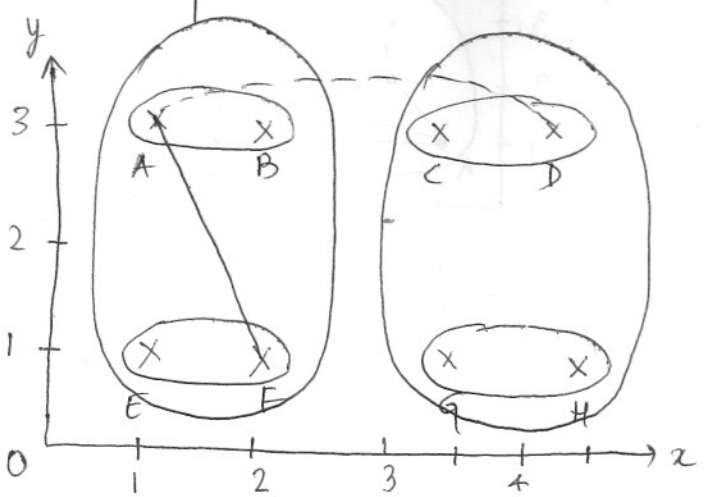
<iv> Similarity $= \dfrac{|AB| + |AC| + |AD| + |BC| + |BD| + |CD|}{6}$

$= \dfrac{1.414 + 4 + 5.099 + 3.162 + 4 + 1.414}{6} = \boxed{3.182}$ Ans.

2. (b) <i> Single link clustering



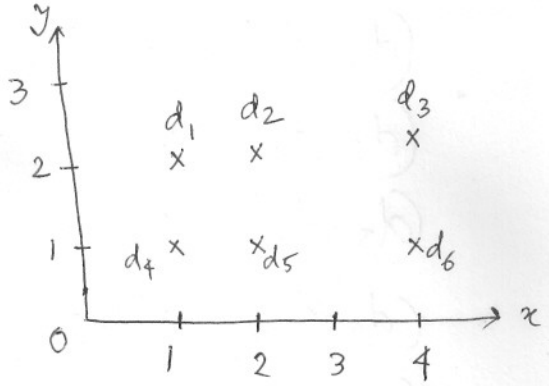
2. (b) <ii> Complete link clustering



Explanations for <i> and <ii>.

- The ellipses correspond to successive clustering stages.
- <i> The single-link similarity of the 2 upper 2-point clusters is the similarity of B and C (solid line), which is greater than the single-link similarity of the 2 left 2-point clusters (dashed line).
- <ii> The complete link similarity of the 2 upper 2-point clusters is the similarity of A and D (dashed line), which is smaller than the complete-link similarity of the 2 left 2-pt. clusters (solid line).

3.



d_1: (1, 2)
d_2: (2, 2)
d_3: (4, 2)
d_4: (1, 1)
d_5: (2, 1)
d_6: (4, 1)

(a) Initial seeds d_2, d_5
(2,2) (2,1)

Distance $(d_1, d_2) = 1$
$(d_3, d_2) = 2$
$(d_4, d_2) = 1.414$
$(d_5, d_2) = 1$
$(d_6, d_2) = 2.236$
$(d_2, d_2) = 0$

$(d_1, d_5) = 1.414$
$(d_2, d_5) = 1$
$(d_3, d_5) = 2.236$
$(d_4, d_5) = 1$
$(d_5, d_5) = 0$
$(d_6, d_5) = 2$

$d_1 \rightarrow$ (assigned to) d_2, $d_2 \rightarrow d_2$, $d_3 \rightarrow d_2$, $d_4 \rightarrow d_5$, $d_5 \rightarrow d_5$, $d_6 \rightarrow d_5$

∴ After iter 1: $\{d_1, d_2, d_3\}$, $\{d_4, d_5, d_6\}$

New centroid for cluster 1: $\frac{1+2+4}{3} = 2$; $\frac{2+2+2}{3} = 2 \Rightarrow (2,2)$ ↓ d_2
(1, 2) (2,2) (4,2)

cluster 2 $\frac{1+2+4}{3} = 2$; $\frac{1+1+1}{3} = 1 \Rightarrow (2,1)$ ↓ d_5
(1, 1), (2,1), (4, 1)

No change in centroids.

∴ Final clusters: $\boxed{\{d_1, d_2, d_3\} \\ \{d_4, d_5, d_6\}}$ Ans.

~~7~~ 6

3. (b) Distance $(d_1, d_2) = 1$ $(d_1, d_3) = 3$

$(d_2, d_2) = 0$ $(d_2, d_3) = 2$

Seeds: $(d_3, d_2) = 2$ $(d_3, d_3) = 0$

d_2, d_3 $(d_4, d_2) = 1.414$ $(d_4, d_3) = 3.162$

$(d_5, d_2) = 1$ $(d_5, d_3) = 2.236$

$(d_6, d_2) = 2.236$ $(d_6, d_3) = 1$

$d_1 \to d_2$, $d_2 \to d_2$, $d_3 \to d_3$, $d_4 \to d_2$, $d_5 \to d_2$, $d_6 \to d_3$

$\therefore$ After phase 1: $\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}$

New centroids:

Cluster 1 $\frac{1+2+1+2}{4} = 1.5$, $\frac{2+2+1+1}{4} = 1.5$

(1,2), (2,2), (1,1), (2,1) $\mu_1 \Rightarrow (1.5, 1.5)$

Cluster 2 $\frac{4+4}{2} = 4$, $\frac{2+1}{2} = 1.5 \Rightarrow (4, 1.5)$

(4,2), (4,1)

μ_2

$d_1(1,2)$ Distance $(d_1, \mu_1) = 0.707$ $(d_1, \mu_2) = 3.041$

$d_2(2,2)$ $(d_2, \mu_1) = 0.707$ $(d_2, \mu_2) = 2.062$

$d_3(4,2)$ $(d_3, \mu_1) = 2.550$ $(d_3, \mu_2) = 0.5$

$d_4(1,1)$ $(d_4, \mu_1) = 0.707$ $(d_4, \mu_2) = 3.041$

$d_5(2,1)$ $(d_5, \mu_1) = 0.707$ $(d_5, \mu_2) = 2.062$

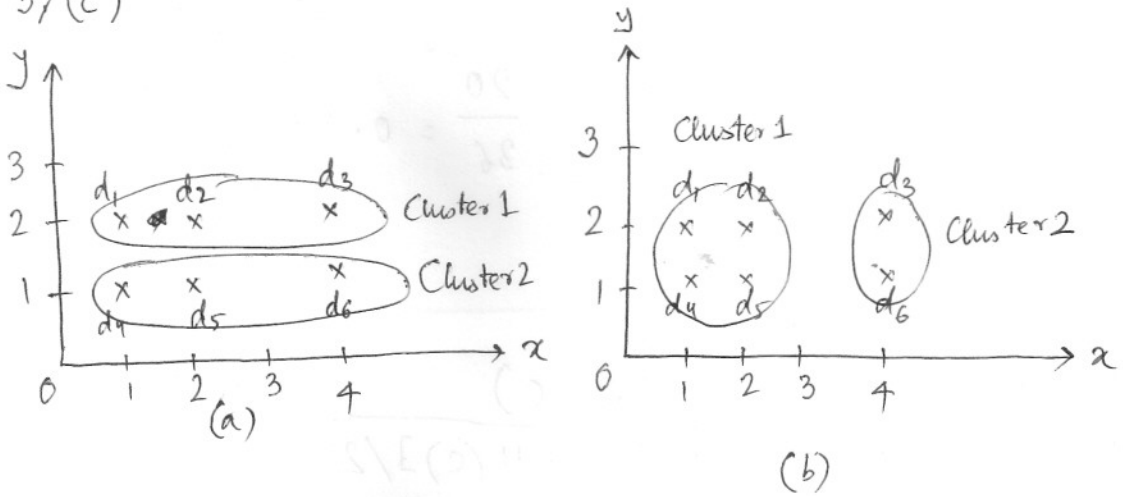
$d_6(4,1)$ $(d_6, \mu_1) = 2.550$ $(d_6, \mu_2) = 0.5$

Assignments: $d_1 \to \mu_1$, $d_2 \to \mu_1$, $d_3 \to \mu_2$, $d_4 \to \mu_1$, $d_5 \to \mu_1$, $d_6 \to \mu_2$

$\therefore$ No reassignment. Final clusters:

$\boxed{\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}}$

3) (c)



---

3) (d) Yes, the 2 clusterings are different.

The clustering for 3.(b) is better, as the clusters are more compact (average), with ~~inter-cl~~ intracluster-to-intercluster distance ratios also being lower than 3.(a).

---

3) (e) The final output of k-means is highly dependent on the choice of initial seeds.

---

4) (a)

$$\text{Purity} = \frac{6+4+6+4}{11+7+11+7} = \frac{20}{36} = 0.556$$

---

4. (b)

$$NMI(\Omega, C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2}$$

$$I(\Omega; C) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log_2 \frac{N\,|\omega_k \cap c_j|}{|\omega_k|\,|c_j|}$$

$N = 36$

$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$

$C = \{c_1, c_2, c_3\}$ [★ △ □]

3 × 4 = 12 terms

cl1 ★	cl2 ★
cl1 △	cl2 △
cl1 □	cl2 □
cl3 ★	cl4 ★
cl3 △	cl4 △
cl3 □	cl4 □

$$I(\Omega; C) = \frac{5}{36}\log_2 \frac{36\times 5}{11\times 18} + \frac{6}{36}\log_2 \frac{36\times 6}{11\times 11}$$

$$+ 0$$

$$+ \frac{4}{36}\log_2 \frac{36\times 4}{7\times 18} + 0 + \frac{2}{36}\log_2 \frac{36\times 3}{7\times 7}$$

$$+ \frac{6}{36}\log_2 \frac{36\times 6}{11\times 18} + \frac{5}{36}\log_2 \frac{36\times 5}{11\times 11} + 0$$

$$+ \frac{2}{36}\log_2 \frac{36\times 3}{7\times 18} + 0 + \frac{4}{36}\log_2 \frac{36\times 4}{7\times 7}$$

$$= \frac{5}{36}\log_2 \frac{10}{11} + \frac{1}{6}\log_2 \frac{216}{121} + \frac{1}{9}\log_2 \frac{8}{7} + \frac{1}{12}\log_2 \frac{108}{49}$$

$$+ \frac{1}{6}\log_2 \frac{12}{11} + \frac{5}{36}\log_2 \frac{180}{121} + \frac{1}{12}\log_2 \frac{6}{7} + \frac{1}{9}\log_2 \frac{144}{49}$$

$$= \frac{5}{36}\log_2\left(\frac{1800}{1331}\right) + \frac{1}{6}\log_2\left(\frac{2592}{1331}\right) + \frac{1}{9}\left(\log_2\left(\frac{1152}{343}\right) + \frac{1}{12}\log_2\left(\frac{648}{343}\right)\right)$$

$$= 0.060 + 0.160 + 0.194 + 0.076$$

$$= 0.490 \checkmark$$

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N}\log_2\frac{|\omega_k|}{N}$$

$$= -\left(\frac{11}{36}\log_2\frac{11}{36} + \frac{7}{36}\log_2\frac{7}{36} + \frac{11}{36}\log_2\frac{11}{36} + \frac{7}{36}\log_2\frac{7}{36}\right)$$

$$= -\left(\frac{11}{36}\log_2\frac{121}{1296} + \frac{7}{36}\log_2\frac{49}{1296}\right)$$

$$= -(-1.045 - 0.919)$$

$$= 1.964$$

$$H(C) = -\sum_j \frac{|c_j|}{N}\log_2\frac{|c_j|}{N}$$

$$= -\left(\frac{18}{36}\log_2\frac{18}{36} + \frac{11}{36}\log_2\frac{11}{36} + \frac{7}{36}\log_2\frac{7}{36}\right)$$

$$= -(-0.5 - 0.523 - 0.459)$$

$$= 1.482$$

$$\therefore NMI(\Omega, C) = \frac{0.490}{\frac{1.964 + 1.482}{2}} = \boxed{0.284} \text{ Ans.}$$

4. (c) Rand Index $= \dfrac{TP + TN}{TP + FP + FN + TN}$

$TP + FP = {}^{11}C_2 + {}^{7}C_2 + {}^{11}C_2 + {}^{7}C_2 =$

$= \dfrac{2(11)(10)}{2} + \dfrac{2(7)(6)}{2} = 152$

$TP = {}^{5}C_2 + {}^{6}C_2 + {}^{4}C_2 + {}^{3}C_2 + {}^{6}C_2 + {}^{5}C_2 + {}^{4}C_2 + {}^{3}C_2$

$= 10 + 15 + 6 + 3 + 15 + 10 + 6 + 3$

$= 68$

$\therefore FP = 152 - 68 = 84$

$TN + FN = {}^{N}C_2 - (TP + FP) = {}^{36}C_2 - 152$

$= 478$

$FN = (5 \times 4) + (5 \times 6) + (5 \times 3) + (4 \times 6) + (4 \times 3) + (6 \times 3)$

~~$+ (6 \times 0) +$~~ (6×5) ~~$+ (6 \times 0) +$~~ $+ (3 \times 4)$

$= 20 + 30 + 15 + 24 + 12 + 18 + 30 + 12$

$= 161$

$\therefore TN = 478 - 161 = 317$

$\therefore RI = \dfrac{68 + 317}{152 + 478} = \dfrac{385}{630} = \boxed{0.611}$ Ans.

4) (d) $\text{Precision} = \frac{68}{152} = 0.447$

$$\text{Recall} = \frac{68}{68+161} = 0.297$$

$$\therefore \text{F-Measure} = \frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{2 \times 0.447 \times 0.297}{0.447 + 0.297}$$

$$= \boxed{0.357} \quad \text{Ans.}$$

5.) (a) Rank = 2 [Row 1 + Row 2 = Row 3; So only 2 linearly independent rows]

---

5. (b) satisfying criterion for an eigenvalue is : $C.\bar{x} = \lambda.\bar{x}$

Let $\bar{x} = (x_1, x_2)^T$ be an eigenvector and $\lambda = 2$ (given) be the eigenvalue.

$$\Rightarrow \begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \cdot \bar{x} = 2\bar{x} \Rightarrow 6x_1 - 2x_2 = 2x_1 \text{ and}$$

$$4x_1 - 0x_2 = 2x_2 .$$

Since the equations are consistent, $\lambda = 2$ is an eigenvalue.

Proved.

Corresponding eigenvector : $\boxed{\bar{x} = (1, 2)}$ Ans.

($2x_1 = x_2$ from above)

---

5. (c)

$$C = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

$$CC^T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad \text{Ans.}$$

- Diagonal elements represent term frequencies (corresponding) in the corpus. Ans.

---

5. (d) The $(i,j)^{th}$ entry in C^TC is the no. of terms that documents i and j have in common.

---

5. (e) Precision generally decreases.
Recall generally increases.

5.(f) Frobenius norm of $X (= C - C_k)$, where

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \text{ and } C_{\substack{k \\ \downarrow \\ 2}} = \begin{pmatrix} -1.62 & -0.60 & -0.44 \\ -0.46 & 0.84 & 0.30 \\ 0 & 0 & 0 \end{pmatrix},$$

is given by $\|X\|_F = \sqrt{\sum_{i=1}^{M} \sum_{i=1}^{N} x_{ij}^2}$

$$X = \begin{pmatrix} 2.62 & 0.60 & 1.44 \\ 0.46 & 0.16 & -0.30 \\ 1 & 1 & 0 \end{pmatrix}$$

$\therefore \|X\|_F = \sqrt{11.625} = \boxed{3.410}$ Ans.

5.(g) Synonymy, Semantic relatedness.

---

6)

Term frequencies:

word	query q	d_1	d_2
CDs	2	2	0
cheap	3	2	1
DVDs	1	0	1
extremely	1	0	0
software	0	1	0
thrills	0	0	1

(14)

For $1.0 \times \vec{q} + 0.75 \times \vec{d} + 1 - 0.25 \times \vec{d_2}$, we get:

$(3.5 \quad 4.25 \quad 0.75 \quad 1 \quad 0.75 \quad -0.25)^T$ or

$(7/2 \quad 17/4 \quad 3/4 \quad 1 \quad 3/4 \quad -1/4)^T$.

Negative weights are set to 0.

∴ The Rocchio vector is:

$\boxed{(3.5 \quad 4.25 \quad 0.75 \quad 1 \quad 0.75 \quad 0)^T}$. Ans.

[Any permutation of the above vector is fine]

---

7) (a)

$$RSV_d = \sum_{t \in q} \left(\log_{10} \left[\frac{N}{df_t} \right] \times \frac{(k_1+1)\, tf_{t,d}}{k_1((1-b) + b(Ld/Lavg)) + tf_{t,d}} \right)$$

$k_1 = 1$ $\quad N = 4$

$b = 0.5$ $\quad L_{avg} = \frac{9+6+8+6}{4} = 7.25$

q : obama health plan

$$RSV_{d_1} = \log_{10} \frac{4}{4} \times \frac{2 \times 1}{1(0.5 + 0.5(9/7.25)) + 1}$$

$$+ \log_{10} \frac{4}{2} \times \frac{2 \times 2}{1(0.5 + 0.5(9/7.25)) + 2}$$

$$+ \log_{10} \frac{4}{2} \times \frac{2 \times 0}{1(0.5 + 0.5(9/7.25)) + 0}$$

$$= 0.386$$

$$RSV_{d_2} = \log_{10}\frac{4}{4} \times \frac{2\times 1}{1(0.5+0.5(6/7.25))+1} + \log_{10}\frac{4}{2} \times \frac{2\times 0}{1(0.5+0.5(6/7.25))+0} + \log_{10}\frac{4}{2} \times \frac{2\times 1}{1(0.5+0.5(6/7.25))+1}$$

$$= 0.315$$

$$RSV_{d_3} = \log_{10}\frac{4}{4} \times \frac{2\times 1}{1(0.5+0.5(8/7.25))+1} + \log_{10}\frac{4}{2} \times \frac{2\times 0}{1(0.5+0.5(8/7.25))+1} + \log_{10}\frac{4}{2} \times \frac{2\times 1}{1(0.5+0.5(8/7.25))+1}$$

$$= 0.293$$

$$RSV_{d_4} = \log_{10}\frac{4}{4} \times \frac{2\times 2}{1(0.5+0.5(6/7.25))+2} + \log_{10}\frac{4}{2} \times \frac{2\times 1}{1(0.5+0.5(6/7.25))+1} + \log_{10}\frac{4}{2} \times \frac{2\times 0}{1(0.5+0.5(6/7.25))+0}$$

$$= 0.315$$

$\therefore$ Reqd. ranking :

$$\boxed{d_1 > d_2 = d_4 > d_3}$$ Ans.

7) (b) 'k₁' is the tuning parameter controlling the document term frequency scaling.

---

7) (c) 'b' is the tuning parameter controlling the scaling by document length.

---

7) (d) When $k_1 = 0$,

$$RSV_d = \sum_{t \in q} \left(\log_{10} \left[\frac{N}{df_t} \right] \times \frac{tf_{t,d}}{tf_{t,d}} \right) = \sum_{t \in q} \log_{10} \frac{N}{df_t}$$

~~So documents are ranked simply in the descending order of the sum of the IDFs of the query terms.~~

The ranking factor simply becomes the sum of query term IDFs. Thus, we cannot get any ranking as the (t, d) factor is lost.

7.(e) We will need to use a weighting for query terms.

Exact formula (need not provide):

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{df_t} \right] \times \frac{(k_1+1)\, tf_{t,d}}{k_1((1-b) + b\left(\frac{L_d}{L_{avg}}\right)) + tf_{t,d}} \times \frac{(k_3+1)\, tf_{t,q}}{k_3 + tf_{t,q}}$$

where $tf_{t,q}$ is the frequency of term t in query q

k_3 is the tuning parameter controlling term frequency scaling of the query.

---

8) (a) Query q : cats and dogs

$$P(q|d_1) = \left[0.25 \times \frac{1}{7} + 0.75 \times \frac{7}{30}\right] \times \left[0.25 \times \frac{1}{7} + 0.75 \times \frac{4}{30}\right]$$

$$\times \left[0.25 \times \frac{1}{7} + 0.75 \times \frac{5}{30}\right]$$

$$= 0.211 \times 0.136 \times 0.161 = 4.62 \times 10^{-3}$$

$$P(q|d_2) = \left[0.25 \times \frac{2}{9} + 0.75 \times \frac{7}{30}\right] \times \left[0.25 \times \frac{1}{9} + 0.75 \times \frac{4}{30}\right]$$

$$\times \left[0.25 \times \frac{1}{9} + 0.75 \times \frac{5}{30}\right]$$

$$= 0.231 \times 0.128 \times 0.153 = 4.52 \times 10^{-3}$$

$$P(q|d_3) = \left[0.25 \times \frac{3}{6} + 0.75 \times \frac{7}{30}\right] \times \left[0.25 \times \frac{1}{6} + 0.75 \times \frac{4}{30}\right]$$

$$\times \left[0.25 \times \frac{1}{6} + 0.75 \times \frac{5}{30}\right]$$

$$= 0.300 \times 0.142 \times 0.167 = 7.11 \times 10^{-3}$$

$$P(q|d_4) = \left[0.25 \times \frac{1}{8} + 0.75 \times \frac{7}{30}\right] \times \left[0.25 \times \frac{1}{8} + 0.75 \times \frac{4}{30}\right]$$

$$\times \left[0.25 \times \frac{2}{8} + 0.75 \times \frac{5}{30}\right]$$

$$= 0.206 \times 0.131 \times 0.188 = 5.07 \times 10^{-3}$$

$\therefore$ Required ranking :

$$\boxed{d_3 > d_4 > d_1 > d_2}$$ Ans.

8) (b) Bayesian updating process. $\alpha = 0.5$.

$$\hat{P}(t|d) = \frac{tf_{t,d} + \alpha \hat{P}(t|M_c)}{L_d + \alpha}$$

d_1 $L_1 = 7$

d_2 $L_2 = 9$

d_3 $L_3 = 6$

d_4 $L_4 = 8$

Query q : cats and dogs

$$P(q|d_1) = \left(\frac{1 + 0.5\left(\frac{7}{30}\right)}{7+0.5}\right) \times \left(\frac{1 + 0.5\left(\frac{4}{30}\right)}{7+0.5}\right) \times \left(\frac{1+0.5\left(\frac{5}{30}\right)}{7+0.5}\right)$$

$$= 0.149 \times 0.142 \times 0.144 = 3.047 \times 10^{-3}$$

$$P(q|d_2) = \frac{2 + 0.5\left(\frac{7}{30}\right)}{9+0.5} \times \frac{1+0.5\left(\frac{4}{30}\right)}{9+0.5} \times \frac{1+0.5\left(\frac{5}{30}\right)}{9+0.5}$$

$$= 0.223 \times 0.112 \times 0.114 = 2.847 \times 10^{-3}$$

$$P(q|d_3) = \frac{3+0.5\left(\frac{7}{30}\right)}{6+0.5} \times \frac{1+0.5\left(\frac{4}{30}\right)}{6+0.5} \times \frac{1+0.5\left(\frac{5}{30}\right)}{6+0.5}$$

$$= 0.479 \times 0.164 \times 0.167 = 1.312 \times 10^{-2}$$

$$P(q|d_4) = \frac{1+0.5\left(\frac{7}{30}\right)}{8+0.5} \times \frac{1+0.5\left(\frac{4}{30}\right)}{8+0.5} \times \frac{2+0.5\left(\frac{5}{30}\right)}{8+0.5}$$

$$= 0.131 \times 0.125 \times 0.245 = 4.012 \times 10^{-3}$$

$\therefore$ Reqd. ranking : $\boxed{d_3 > d_4 > d_1 > d_2}$ Ans.

9. (a) c = sports; $\bar{c}$ = $\overline{\text{sports}}$. No. of docs in $c = 3$. No. of docs in $\bar{c} = 2$

$\hat{P}(c) = 3/5$, $\hat{P}(\bar{c}) = 2/5$ (Prior estimates)

$B = 2 \rightarrow$ there are 2 cases to consider; occurrence & non-occurrence

$$P(\text{football} \mid c) = \frac{3+1}{3+2} = 0.8 \quad \rightarrow \text{Laplace smoothing}$$

$$P(\text{football} \mid \bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$P(\text{cricket} \mid c) = \frac{1+1}{3+2} = 0.4$$

$$P(\text{cricket} \mid \bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$P(\text{termite} \mid c) = \frac{0+1}{3+2} = 0.2$$

$$P(\text{termite} \mid \bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$P(\text{grasshopper} \mid c) = \frac{0+1}{3+2} = 0.2$$

$$P(\text{grasshopper} \mid \bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$P(\text{hockey} \mid c) = \frac{1+1}{3+2} = 0.4$$

$$P(\text{hockey} \mid \bar{c}) = \frac{0+1}{2+2} = 0.25$$

$$P(\text{goal} \mid c) = \frac{1+1}{3+2} = 0.4$$

$$P(\text{goal} \mid \bar{c}) = \frac{0+1}{2+2} = 0.25$$

Terms:

1. football
2. cricket
3. termite
4. grasshopper
5. hockey
6. goal
7. obama
8. romney

(20)

$$P(\text{obama}|c) = \frac{0+1}{3+2} = 0.2$$

$$P(\text{obama}|\bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$P(\text{romney}|c) = \frac{0+1}{3+2} = 0.2$$

$$P(\text{romney}|\bar{c}) = \frac{1+1}{2+2} = 0.5$$

$$\therefore \hat{P}(c|d_6) = \overset{3/5\times}{0.8} \times 0.4 \times 0.4 \times 0.2 \times (1-0.2) \times (1-0.4) \times (1-0.2) \times (1-0.2)$$

$$= 4.719 \times 10^{-3}$$

and, $\hat{P}(\bar{c}|d_6) = \frac{2}{5} \times 0.5 \times 0.5 \times 0.25 \times 0.5 \times (1-0.5) \times (1-0.25) \times (1-0.5) \times (1-0.5)$

$$= 1.172 \times 10^{-3}$$

$\therefore$ The more probable class for d_6 would be **sports**.

Ans.

---

9) (b) $N = 20+30+25+100 = \cancel{225}\ 175$.

$$\chi^2(D, \text{football}, \text{sports}) = \frac{(N_{00}-E_{00})^2}{E_{00}} + \frac{(N_{01}-E_{01})^2}{E_{01}}$$

$$+ \frac{(N_{10}-E_{10})^2}{E_{10}} + \frac{(N_{11}-E_{11})^2}{E_{11}}$$

$$E_{00} = N \times P(\bar{t}) \times P(\bar{c}) = 175 \times \frac{25+100}{175} \times \frac{30+100}{175}$$

$$= 92.857$$

$$E_{01} = N \times P(\bar{t}) \times P(c) = 175 \times \frac{25+100}{175} \times \frac{20+25}{175}$$

$$= 32.143$$

$$E_{10} = N \times P(t) \times P(\bar{c}) = 175 \times \frac{20+30}{175} \times \frac{30+100}{175}$$

$$= 37.143$$

$$E_{11} = N \times P(t) \times P(c) = 175 \times \frac{20+30}{175} \times \frac{20+25}{175}$$

$$= 12.857$$

$$\therefore \chi^2 = \frac{(100-92.857)^2}{92.857} + \frac{(25-32.143)^2}{32.143} + \frac{(30-37.143)^2}{37.143} + \frac{(20-12.857)^2}{12.857}$$

$$= 0.549 + 1.589 + 1.374 + 3.968$$

$$= \boxed{7.480} \text{ Ans.}$$

$\because$ 7.480, > 2.71, we can reject the hypothesis that the occurrence of football is independent of the occurrence of sports. Ans.

(22)

10) (a) N = 6 ~~term weights~~

t_1 = football, t_2 = cricket, t_3 = termite, t_4 = grasshopper,
t_5 = hockey, t_6 = goal, t_7 = obama, t_8 = romney

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
d_1	2	1	0	0	0	0	0	0
d_1 wt.	0.103	0.301	0	0	0	0	0	0
d_2	0	1	1	1	0	0	0	0
d_2 wt.	0	0.301	0.477	0.778	0	0	0	0
d_3	2	0	0	0	1	0	0	0
d_3 wt.	0.103	0	0	0	0.477	0	0	0
d_4	1	0	0	0	0	1	0	0
d_4 wt.	0.079	0	0	0	0	0.778	0	0
d_5	1	0	0	0	0	0	1	1
d_5 wt.	0.079	0	0	0	0	0	0.778	0.778
d_6	1	1	1	0	1	0	0	0
d_6 wt.	0.079	0.301	0.477	0	0.477	0	0	0
df_t	5	3	2	1	2	1	1	1
$\log_{10}(N/df_t)$	0.079	0.301	0.477	0.778	0.477	0.778	0.778	0.778

~~$|\vec{d_1} - \vec{d_6}| = \sqrt{(0.103 - 0.079}$~~

$|\vec{d_1} - \vec{d_6}| = \sqrt{(5.76\times10^{-4}) + 0 + 0.228 + 0.228} = 0.675$

$|\vec{d_2} - \vec{d_6}| = \sqrt{(6.241\times10^{-3}) + 0 + 0 + 0.605 + 0.228} = 0.916$

$|\vec{d_3} - \vec{d_6}| = \sqrt{(5.76\times10^{-4}) + 0.091 + 0.228} = \boxed{0.565}$

$|\vec{d_4} - \vec{d_6}| = \sqrt{0.091 + 0.228 + 0.228 + 0.605} = 1.073$

$|\vec{d_5} - \vec{d_6}| = \sqrt{0 + 0.091 + 0.228 + 0.228 + 0.605 + 0.605} = 1.326$

$\therefore$ d_6 would be assigned to the class of d_3, i.e. Sports Am.

10) (b) $\vec{a} = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}$, $\vec{b} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$, $\vec{c} = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$, $\vec{x} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

<i> Dot product : $\vec{c}$ ($\vec{a}$ 0.05, $\vec{b}$ 0.16, $\vec{c}$ 0.28)

<ii> Cosine : $\vec{b}$ ($\vec{a}$ 0.9805, $\vec{b}$ 1.0, $\vec{c}$ 0.9899)

<iii> Euclidean : $\vec{a}$ ($\vec{a}$ 0.1118, $\vec{b}$ 0.2828, $\vec{c}$ 0.7211)

---

10) (c) Infinite

---

10) (d) k-Nearest Neighbour classifier.

(Naïve Bayes, Rocchio, SVM → ~~both~~ all linear)