Information Retrieval (CS60092)
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

End Semester Examination

Time: 3 hours

Attempt as much as you can.
Make suitable assumptions if needed.
Solution steps / answers should be supported by proper arguments.

---

1) Consider the following matrix representing **distance** between six documents:

Document	A	B	C	D	E	F
A	0	662	877	255	412	996
B	662	0	295	468	268	400
C	877	295	0	754	564	138
D	255	468	754	0	219	869
E	412	268	564	219	0	669
F	996	400	138	869	669	0

Compute hierarchical single-linkage clustering of these six documents. Clearly show the matrices at each step of building the dendrogram.
(No marks will be given for showing only the Final Dendrogram)

[10]

Ans –

The nearest pair of document is C and F, at distance 138. These are merged into a single cluster called "C/F".
Then we compute the distance from this new compound object to all other objects. In single link clustering the rule is that the distance from the compound object to another object is equal to the shortest distance from any member of the cluster to the outside object.
So the distance from "C/F" to E is chosen to be 564, which is the distance from C to E, and so on.

After merging C with F, we obtain the following matrix:

Document	A	B	C/F	D	E
A	0	662	877	255	412

B	**662**	**0**	**295**	**468**	**268**
C/F	**877**	**295**	**0**	**754**	**564**
D	**255**	**468**	**754**	**0**	**219**
E	**412**	**268**	**564**	**219**	**0**

min d(i,j) = d(D,E) = 219 => merge D and E into a new cluster called “D/E”

After merging D with E, we obtain the following matrix:

Document	**A**	**B**	**C/F**	**D/E**
A	**0**	**662**	**877**	**255**
B	**662**	**0**	**295**	**268**
C/F	**877**	**295**	**0**	**564**
D/E	**255**	**268**	**564**	**0**

min d(i,j) = d(A,D/E) = 255 => merge A and D/E into a new cluster called A/D/E

Document	**A/D/E**	**B**	**C/F**
A/D/E	**0**	**268**	**564**
B	**268**	**0**	**295**
C/F	**564**	**295**	**0**

min d(i,j) = d(A/D/E,B) = 268 => merge A/D/E and B into a new cluster called A/D/E/B

Document	**A/D/E/B**	**C/F**
A/D/E/B	**0**	**295**
C/F	**295**	**0**

Finally, we merge the last two clusters at distance 295.

2) Consider the problem of learning to classify a name as being Food or Beverage. Assume the following training set:

Document	**Class**
Cherry Pie Chocolate	Food
Chicken Wings Crispy	Food
Cream Soda Water	Beverage
Orange Soda	Beverage

Train a Multinomial Naive Bayes Classifier on the above data. Calculate the multinomial parameters (Priors and Conditional Probabilities). Use *Laplace Smoothing* for calculation of conditional probabilities.

What does this classifier predict about the class of the following test document: **"Chocolate Cream Soda "** ? Assume *positional independence* of terms.

[7 + 3 = 10]

Ans –

We denote the two classes Food and Beverages by F and B respectively. There are 10 distinct terms. We use unigram based naïve bayes classifier (multinomial model) with laplace smoothing for classification.

Term (t)	P(t|F) [Raw]	P(t|B) [Raw]	P(t|F) [Smoothed]	P(t|B) [Smoothed]
Chocolate	1/6	0/5	(1+1)/(6+10) = 2/16	(0+1)/(5+10) = 1/15
Cream	0/6	1/5	(0+1)/(6+10) = 1/16	(1+1)/(5+10) = 2/15
Soda	0/6	2/5	(0+1)/(6+10) = 1/16	(2+1)/(5+10) = 3/15

P(F|Chocolate cream soda) α P(F) * P(Chocolate|F) * P(Cream|F) *P(soda|F)

= (2/4) *(2/16) * (1/16) * (1/16)

= 4/(4 * 16 * 16 * 16).

P(B|Chocolate cream soda) α P(B) * P(Chocolate|B) * P(Cream|B) *P(soda|B)

= (2/4) *(1/15) * (2/15) * (3/15)

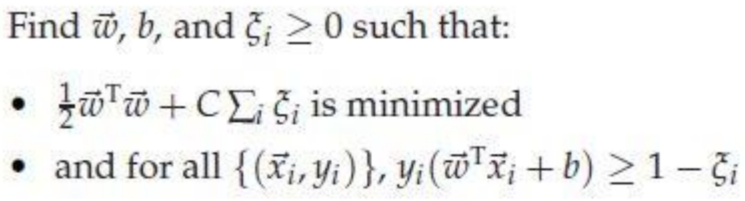
= 12/(4 * 15 * 15 * 15).

As P(B|Chocolate cream soda) > P(F|Chocolate cream soda), the predicted class for the test document is "Beverage"

3)

a) Write and explain the primal formulation of the optimization problem for building a soft margin SVM.

Ans –



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

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

b) Derive the equation of the hard margin SVM classifier for the following set of labeled points.

Point	x1	x2	Class
P1	10	3	+1
P1	8	2	+1
P2	4	0	+1
P3	4	2	-1
P4	2	1	-1

Ans –

Points from the positive classes are on a straight line x1 = 2*x2+4

Points from the negative classes are on a straight line x1 = 2*x2

Hence equation of the hard margin SVM classifier will be x1 = 2*x2 + 2

c) Assume that few more points are added in the following order. What should be the equation of the SVM classifier after addition of each of these points?

Point	x1	x2	Class
P5	6	0	+1
P6	0	0	-1
P7	4	1	-1

[5 + 5 + 5 = 15]

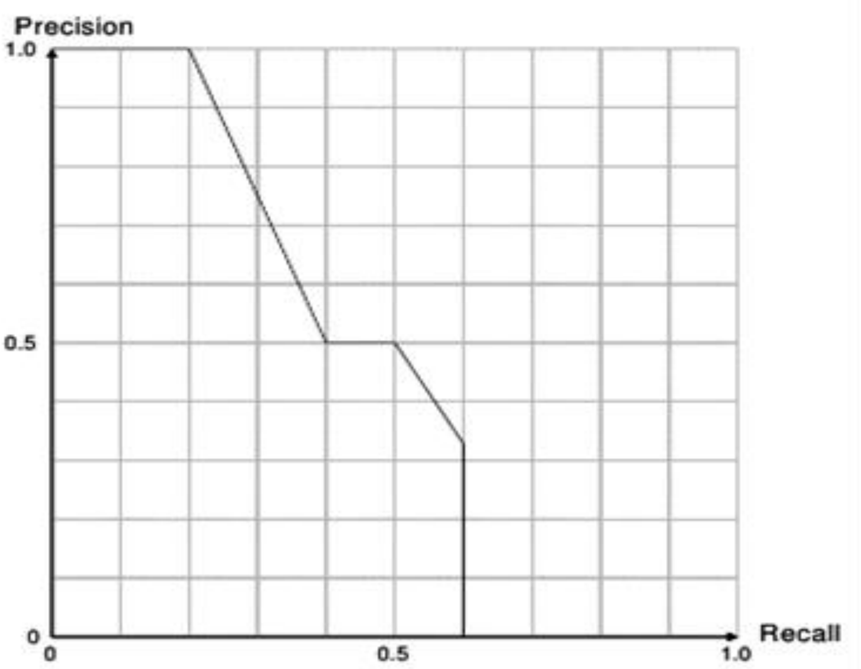
Ans –

P5 falls on the line x1 = 2*x2+4 and the classifier will not change.

P6 will be correctly classified as a negative example and will not fall inside the margin. So the classifier will not change.

P7 will be correctly classified as a positive example and will fall inside the margin. So the classifier will change. The new classifier will be: x1 = 2*x2+3.

4) A document retrieval system produced the following interpolated precision-recall curve) on a particular query (based on 20 results):



You know that there are ***ten*** relevant documents.

a) What is the precision after the system has retrieved three relevant documents?

Ans – 0.75

b) Going down the hit list, you discovered that you have retrieved *n* documents, and all of them are relevant. What is the maximum possible value of *n*?

Ans – 2

c) What are the positions in the ranked list of 20 results that represent relevant documents?

Ans – 1, 2, 4, 8, 10, 18

d) Suppose the relevance label for the relevant documents is 1, and relevance label for the non-relevant documents is 0. Find the NDCG@20 of the result set.

[2 + 2 + 6 + 5 = 15]

Ans –

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log_2(1+m)},$$

Z_{kj} = Best possible DCG value for this query.

= $1/\log_2 2 + 1/\log_2 3 + \ldots + 1/\log_2 11$ (since there are 10 relevant documents)

= 4.5436

DCG for the result set = $1/\log_2 2 + 1/\log_2 3 + 1/\log_2 5 + 1/\log_2 9 + 1/\log_2 11 + 1/\log_2 19$

= 1 + 1/1.5849 + 1/2.3219 + 1/3.1699 + 1/3.4594 + 1/4.24792

= 2.9015

NDCG = 2.9015/4.5436 = 0.6386

5) Consider the following documents:

D1: English Channel Atlantic

D2: National Geography Channel English

D3: Doordarshan National English News

Using unigram language model, rank the above documents for the query

“National News Channel English”.

To compute the model probabilities, combine MLE estimates from documents and the collection giving equal importance to both.

[10]

Ans –

Using linear interpolation smoothing with \alpha = 0.5.

Term	**P(t|D1)**	**P(t|D2)**	**P(t|D3)**	**P(t|C)**
National	0	1/4	¼	2/11
News	0	0	¼	1/11
Channel	1/3	1/4	0	2/11
English	1/3	1/4	¼	3/11

Query (q) = “national news channel English”.

P(q|D1) = (0 + 2/11) / 2 * (0 + 1/11) / 2 * (1/3 + 2/11) / 2 * (1/3 + 3/11) / 2

= $3.225 * 10^{-4}$

P(q|D2) = (1/4 + 2/11) / 2 * (0 + 1/11) / 2 * (1/4 + 2/11) / 2 * (1/4 + 3/11) / 2

= $5.538 * 10^{-4}$

P(q|D3) = (1/4 + 2/11) / 2 * (1/4 + 1/11) / 2 * (0 + 2/11) / 2 * (1/4 + 3/11) / 2

= $8.744 * 10^{-4}$

Hence, the ranking is: D3 > D2 > D1.

6)

a) Assuming Zipf’s law with a corpus independent constant A = 0.1, what is the fraction of words that appear more than 5 times in any fixed corpus of W words?

Ans –

Rank of the most frequent word with frequency 5 = WA/5 = W/50.
Number of words with frequency more than 5 is floor(W/50)
OR
About W/50 words in the collection occur more than 5 times.

b) For a search result set, value of reciprocal rank (RR) is 0.125. What are the maximum and minimum possible values of average precision at position 10 (AP@10) for the result set?

Ans –

First relevant result is at position 8.
Best case: Relevant document at both positions 9 and 10.
Maximum possible AP@10 = (1/8 + 2/9 + 3/10)/10 = 0.064.

Worst case: Non-relevant documents at both positions 9 and 10.
Minimum possible AP@10 = 1/(8*10) = 0.0125

c) Suppose that *C* is a binary term-document incidence matrix. What do the entries of C^TC represent? Explain your answer properly.

[3 + 3 + 4 = 10]

Ans –
(i,j)th entry of CTC denotes the number of number of common terms in document i and j.

7) Consider the following term document matrix C.

Terms	D1	D2	D3	D4	D5	D6
Ship	1	0	1	0	0	0
Boat	0	1	0	0	0	0
Ocean	1	1	0	0	0	0
Voyage	1	0	0	1	1	0
Trip	0	0	0	1	0	1

a) Suppose vector space model is used to represent the documents. Vector dimensions are filled with raw frequency counts of the corresponding terms. According to this representation, what is the similarity between the documents D2 and D3?

b) C is decomposed as $C = U\Sigma V^T$. The matrices U, ∑ and V are given below.

U =

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0	0.73
ocean	−0.48	−0.51	−0.37	0	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

∑ =

$$\begin{matrix} 2.16 & 0 & 0 & 0 & 0 \\ 0 & 1.59 & 0 & 0 & 0 \end{matrix}$$

0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

V^T =

	d1	d2	d3	d4	d5	d6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0	0	0.58	0	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

i) Suppose a low rank approximation of C is obtained as C_2 by keeping the *most* important two terms. According to $C_{2,}$ what is the similarity between documents D2 and D3?

C_2 =

[0.8511 0.5189 0.2807 0.1272 0.2087 -0.0815
0.3628 0.3567 0.1559 -0.2042 -0.0228 -0.1814
1.0128 0.7201 0.3614 -0.0443 0.1637 -0.2081
0.9726 0.1284 0.1967 1.0310 0.6214 0.4096
0.1215 -0.3905 -0.0840 0.9038 0.4127 0.4911]

Sim(D2,D3)
= < [0.52 0.36 0.72 0.13 -0.39], [0.28 0.16 0.36 0.20 -0.08]>
= .52 (Similarity is calculated using inner product)
= .9417 (according to cosine similarity)

ii) Suppose another low rank approximation of C is obtained as C'_2 by keeping the *least* important two terms. According to $C'_{2,}$ what is the similarity between documents D2 and D3?

C'_2 =

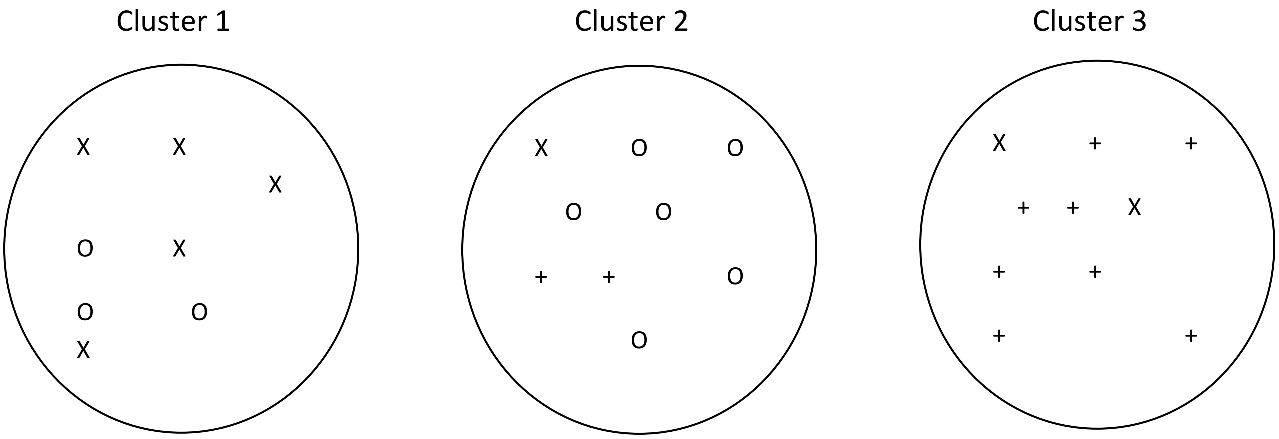
[0.8511 0.5189 0.2807 0.1272 0.2087 -0.0815
0.3628 0.3567 0.1559 -0.2042 -0.0228 -0.1814
1.0128 0.7201 0.3614 -0.0443 0.1637 -0.2081
0.9726 0.1284 0.1967 1.0310 0.6214 0.4096
0.1215 -0.3905 -0.0840 0.9038 0.4127 0.4911]

Sim(D2,D3) =
= < [0.03 0.08 -0.07 0.02 -0.01], [0.40 0.18 -0.15 -0.30 0.31]>
= .028 (Similarity is calculated using inner product)
=.3896 (according to cosine similarity)

c) Find out the Eigen Values of the matrix CC^T.
Ans: 4.68, 2.54, 1.63, 1.00, 0.16

[2 + 2 + 2 + 4 = 10]

8) Consider the following figure for clusters found after performing flat clustering (K-Means) on a set of documents. The gold standard for each document is produced by human judges. Each document belongs to one of the three gold standard classes (x, o and +)



Calculate the following quality measures for the above clustering

a) Purity
b) NMI
c) Rand Index
d) F Measure

[2 + 4 + 2 + 2 = 10]

Ans –

a) To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N.

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

Purity = (1/27) * (5+6+8) = 19/27 = 0.703703704

b) NMI is based on defining a confusion matrix N, where the rows correspond to the gold standard classes and the columns correspond to the clusters found.
The member of N, N_{ij} is simply the number of nodes in the class i that appear in the found cluster j. The number of classes is denoted C_A and the number of found clusters is

denoted by C_B. The sum over row i of matrix N_{ij} is denoted N_i; and the sum over column j is denoted N_j.

$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij} log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A} N_{i.} log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{c_B} N_{.j} log\left(\frac{N_{.j}}{N}\right)}$$

NMI = [-2*[5*log(5*27/8*8)+1*log(1*27/8*9)+2*log(2*27/8*10)
+3*log(3*27/9*8)+6*log(6*27/9*6)+2*log(2*27/10*9)
+8*log(8*27/10*10)]] / [8*log(8/27)+9*log(9/27) +10*log(10/27)
+9*log(9/27)+9*log(9/27)+10*log(10/27)]

= 6.1015*2/25.736 = 0.475

c) A true positive (TP) decision assigns two similar documents to the same cluster; a true negative (TN) decision assigns two dissimilar documents to different clusters.
There are two types of errors we can commit. A false positive (FP) decision assigns two dissimilar documents to the same cluster. A false negative (FN) decision assigns two similar documents to different clusters. The Rand Index (RI) measures the percentage of decisions that are correct.

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

We first compute TP +FP.
The three clusters contain 8, 9 and 10 points, respectively, so the total number of “positives” or pairs of documents that are in the same cluster is:

TP + FP = 8C2 + 9C2 + 10C2 = 109

Of these, the x pairs in cluster 1 and 3, the o pairs in cluster 1 and 2, the + pairs in cluster 2 and 3 are true positives.

TP = 5C2 + 2C2 + 3C2 + 6C2 + 8C2 + 2C2 = 58

Thus, FP = 109 – 58 = 51

FN = 5*2 + 5*1 +2*1 + 6*3 + 8*2 = 51

TN = Total – (TP+FP+FN) = 27C2 – (58+51+51) = 351 – 160 = 191

Thus, RI = (58+191)/351 = 249/351 = 0.71

d) P = TP/ (TP+FP) = 58/ (58+51) = 0.53
R =TP/ (TP+FN) = 58/ (58+51) = 0.53

F Measure = 2PR/(P+R) = 0.53.