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

End-Semester Examination, Autumn 2012

Answer as many questions as you can. *Time: 3 hours*
Use of calculator is allowed. *Maximum Marks: 100*
State any assumptions made clearly.

Q. 1> We wish to build an SVM classifier that categorizes a point into Class A or B based on 3 data points: (1, 1) (Class A), (2, 0) (Class A) and (2, 3) (Class B).

(a) Find the optimal separating hyperplane $\vec{w}$.

(b) Find the margin ρ.

(c) Plot the points on a small graph and depict the separator geometrically. Mark points, lines and distances clearly.

(d) For a soft margin classification problem (taking misclassifications into account) with an SVM, state the primal formulation of the problem that includes the slack variables ξ_i-s and the regularization parameter *C*. State only the problem formulation (not the solution) providing only the expressions (do not describe). Mention the constraint(s).

(e) State the dual formulation of the above problem that uses Lagrange multipliers α_i-s. State only the problem formulation (not the solution) providing only the expressions (do not describe). Mention the constraint(s).

(f) Name the popular class of functions that helps an SVM find a linear separating hyperplane in a higher-dimensional space, when the data in the original dimensions is not linearly separable.

(3 + 1 + 1 + 2 + 2 + 1 = 10)

Q. 2> (a) Consider the 4 points A (1, 1), B (2, 2), C (5, 1) and D (6, 2). Assume A and B are in Cluster 1 and C and D are in Cluster 2. For each of the 4 cases below, provide a graphical representation of the points, clusters and the inter-point distances considered while computing inter-cluster similarity. Also compute this distance (similarity) between the 2 clusters assuming:
<i> Single link clustering
<ii> Complete link clustering
<iii> Centroid-based clustering
<iv> Group average-based clustering

(b) Consider the 8 points A (1, 3), B (2, 3), C (3.5, 3), D (4.5, 3), E (1, 1), F (2, 1), G (3.5, 1) and H (4.5, 1). We wish to divide them into 2 clusters (finally). Plot the points on a graph and draw successive stages of the clustering process assuming (give separate diagrams for <i> and <ii>)
<i> Single link clustering
<ii> Complete link clustering.
Mark appropriate inter-point distances used in the clustering decision process. Describe the process happening in both cases in brief. **(6 (1 + 1 + 2 + 2) + 4 (2 + 2) = 10)**

Q. 3> Consider the documents represented as points on the *x*-*y* plane: d_1 (1, 2), d_2 (2, 2), d_3 (4, 2), d_4 (1, 1), d_5 (2, 1) and d_6 (4, 1).

(a) Perform *k*-means clustering (till convergence) with d_2 and d_5 as initial seeds.

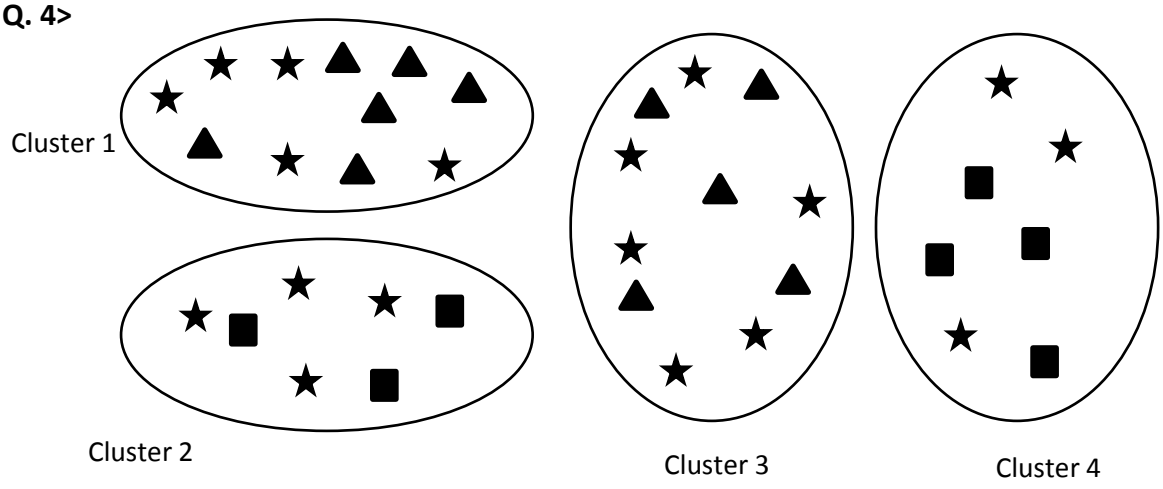
(b) Perform *k*-means clustering (till convergence) with d_2 and d_3 as initial seeds.

(c) Show the points and the clusters obtained in both cases using a rough geometrical plot.

(d) Are the two clusterings different? If yes, which do you think is better? Give reason (qualitative justification, no need to compute any metric).

(e) What disadvantage of *k*-means is visible from these observations? **(3 + 3 + 2 + 1 + 1 = 10)**

Q. 4>



(a) Compute the purity of this clustering scheme.

(b) Compute the NMI of this clustering scheme (log base = 2).

(c) Compute the Rand Index of this clustering scheme.

(d) Compute the *F*-Measure of this clustering scheme ($\beta = 1$). **(2 + 4 + 2 + 2 = 10)**

Q. 5> (a) What is the rank of the 3x3 matrix shown below (Do not show elaborate computations, one quick observation is sufficient)?

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

(b) Show that $\lambda = 2$ is an eigenvalue of

$$C = \begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix}.$$

Find the corresponding eigenvector.

(c) Consider the three terms *tug, war* and *peace* and the three documents d_1: *tug of war*, and d_2: *tug of peace,* and d_3: *war and peace*. Build the term co-occurrence matrix CC^T, where *C* is the term-document incidence matrix. What do the diagonal elements represent?

(d) What do the elements of C^TC represent?

(e) What is the effect of applying LSI on precision and recall?

(f) During the process of applying LSI, a matrix

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

was approximated as the lower rank matrix

$$C_2 = \begin{pmatrix} -1.62 & -0.60 & -0.44 \\ -0.46 & 0.84 & 0.30 \\ 0 & 0 & 0 \end{pmatrix}$$

What is the Frobenius norm of the error of this approximation?

(g) Name one semantic challenge during document processing that LSI can address, but the vector space model cannot. **(1 + 2 + 2 + 1 + 1 + 2 + 1 = 10)**

Q. 6> Suppose that a user's initial query is *cheap cds cheap dvds extremely cheap cds*. The user examines two documents d_1 and d_2. He judges d_1, with the content *cds cheap software cheap cds* relevant and d_2 with content *cheap thrills dvds* non-relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback as in

$$\overrightarrow{q_m} = \alpha\overrightarrow{q_0} + \beta\frac{1}{|D_r|}\sum_{\overrightarrow{d_j}\in D_r}\overrightarrow{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\overrightarrow{d_j}\in D_{nr}}\overrightarrow{d_j}$$

what would the revised query vector be after relevance feedback? Assume α = 1, β = 0.75, γ = 0.25. **(10)**

Q. 7> The BestMatch25 (Okapi BM25) retrieval model is an improvement over BIM. It considers term frequency in documents, and the document length, during ranking. In BM25, the Retrieval Status Value (RSV) of a document *d* is given by

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

where *N* is the no. of documents in the corpus

L_d is the length of the document *d* in words

L_{avg} is the average document length in words

t, q, d, tf, and *df* have their usual meanings

k_1 and *b* are tunable parameters

Now consider the query *obama health plan*. The document collection consists of the following 4 documents only.:

d_1: *obama rejects health allegations about his own bad health*

d_2: *the plan is to visit obama*

d_3: *obama raises concerns with us medical plan reforms*

d_4: *obama states an obama health vision*

(a) Provide a ranking for the documents in response to the query. Show all intermediate steps of the computation. Assume k_1 = 1, *b* = 0.5.

(b) What is the possible role of the parameter k_1?

(c) What is the possible role of the parameter *b*?

(d) What is the interpretation of BM25 when k_1 = 0?

(e) What possible modification (need not give any definite formula) may be needed for this technique if the query is very long, possibly containing multiple occurrences of the same term?

(6 + 1 + 1 + 1 + 1 = 10)

P. T. O.

Q. 8> Let the user's query be *cats and dogs*. The document collection contains only the following four documents:

d_1: *cats are small and so are dogs*
d_2: *cats and dogs may live as long as cats*
d_3: *dogs attack cats, cats and cats*
d_4: *dogs and cats may be friends of dogs*

(a) Rank the documents in response to the query, using the unigram MLE model from the document *and* the collection. Mix the models with λ = 0.25. Show all steps of the computation.

(b) Now consider the Bayesian updating process. Rank the documents accordingly, considering tunable parameter $\alpha = 0.5$.
Hint: The equation to use for Bayesian updating is

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

(5 + 5 = 10)

Q. 9> (a) Following the multivariate Bernoulli classifier, what would be the more probable class for document 6 (see Table below)? Use Laplace smoothing in your computations. Clearly state all the prior estimates and the conditional probabilities involved.

	docID	**words in document**	**in class = *sports*?**
training set	1	*football cricket football*	Yes
	2	*cricket termite grasshopper*	No
	3	*football football hockey*	Yes
	4	*football goal*	Yes
	5	*obama romney football*	No
test set	6	*football cricket hockey termite*	?

(b) Consider the term t = football and the class c = *sports*. The counts of the number of documents in the corpus D with the four possible combinations of indicator values are as follows:

	$e_c = e_{sports} = 1$	$e_c = e_{sports} = 0$
$e_t = e_{football} = 1$	$N_{11} = 20$	$N_{10} = 30$
$e_t = e_{football} = 0$	$N_{01} = 25$	$N_{00} = 100$

The χ^2 test is applied to test the independence of a term and the occurrence of a class. The χ^2 quantity associated with a term-class pair with respect to the collection is given by

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

where e_t, e_c are as defined in the table

N is the total no. of documents in *D*

$N_{e_t e_c}$ are the observed frequencies as in the table

$E_{e_t e_c}$ are the expected frequencies, which are computed as the product of *N*, *P(t)* (or $P(\bar{t})$) and *P(c)* (or $P(\bar{c})$), according as the case may be. For example, E_{10} = *N* x *P(t)* x $P(\bar{c})$.

Compute the χ^2 for *football* with respect to *sports*. Assuming the critical χ^2 value to be 2.71, is the occurrence of *football* independent of the occurrence of *sports*? **(5 + 5 = 10)**

Q. 10> (a) Consider the set of six documents below as your entire collection. Use the TF-IDF weight formula $w_{t,d}$ = (1+ $\log_{10} tf_{t,d}$)$\log_{10}(N/df_t)$, where *N* is the no. of documents in the collection. Compute the unnormalized weight vectors for each of the six documents and use them to classify d_6 using the 1-Nearest Neighbor method.

	docID	**words in document**	**in class = *sports*?**
training set	1	*football cricket football*	Yes
	2	*cricket termite grasshopper*	No
	3	*football football hockey*	Yes
	4	*football goal*	Yes
	5	*obama romney football*	No
test set	6	*football cricket hockey termite*	?

(b) Consider four vectors $\vec{a} = (0.5\ 1.5)^T$, $\vec{b} = (4\ 4)^T$, $\vec{c} = (8\ 6)^T$ *and* $\vec{x} = (2\ 2)^T$. Which of $(\vec{a},\ \vec{b}\ and\ \vec{c})$ is the most similar to according $\vec{x}$ to

<i> Dot product similarity

<ii> Cosine similarity?

<iii> Euclidean distance?

(c) How many hyperplanes separate two linearly separable classes?

(d) Name one non-linear classifier. **(5 + 3 + 1 + 1 = 10)**