Student Name ____________________

Write your answers on the pages; use the back of the pages if needed. Clear hand writing is required; if we cannot read what you write, it would be considered incorrect.

Table of Content

<table>
<thead>
<tr>
<th>Problem</th>
<th>Title</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Text Categorization: LR and SVM</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Learning to Rank</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Clustering</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Collaborative Filtering</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>Federated Search?</td>
<td>20</td>
</tr>
</tbody>
</table>
1. **Text Categorization: LR and SVM [16 points]**

   a) In logistic regression (LR), the decision boundary can be defined as a set of data points which satisfy the equality $P_w(y = 1 | x) = P_w(y = 0 | x)$ where
   
   $$P_w(y = 1 | x) = \frac{1}{1 + \exp(-w^T x)}.$$ 

   Use this property to prove that LR is a linear classifier, i.e., $w^T x = 0$ is the decision boundary in LR. \([10 \text{ points}]\)

   b) In support vector machines (SVM), the margin ($\gamma$) is defined as the perpendicular distance from any support vector ($x^+$ or $x^-$) to the hyperplane. Use a key property of support vectors to prove $\gamma = \frac{1}{\|w\|}$. \([6 \text{ points}]\)

   [Hint: The perpendicular distance from vector $x$ to hyperplane $h$ is:
   
   $$d(x, h) = \frac{y f_{w,b}(x)}{\|w\|} \quad \text{where} \quad y \in \{1, -1\} \quad \text{and} \quad f_{w,b}(x) = w^T x - b$$
   ]

   Answers:
2. Learning to Rank (LETOR) [24 points]

You have learned that the SVM software can be directly used, or extended, for imposing constraints in LETOR for IR. Provide answers for the following questions [6 point each].

a) Describe (in English) the intuition in the pairwise constraints for training rankSVM.

b) Describe (in English) the intuition in the listwise constraints for training SVM-MAP.

c) Provide the precise formula for the pairwise constraints in rankSVM.

d) Provide the precise formula for the listwise constraints in SVM-MAP.

[You do not need to specify the features in representing a q-d pair vector or a ranked list.]

Notation:

- \( D = (d_1, d_2, \ldots, d_N) \), the training documents
- \( Q = (q_1, q_2, \ldots, q_K) \), the training queries
- \( x(q_k, d_i) \in \mathbb{R}^m \), the feature vector representing the pair of \( q_k \in Q \) and \( d_i \in D \)
- \( D_i^{(k)} \equiv \{d_i \in D : d_i \text{ is relevant to } q_k\} \) and \( D_j^{(k)} \equiv \{d_j \in D : d_j \text{ is irrelevant to } q_k\} \)
- \( L = \{l\} \), the set of all possible ranking of the documents in \( D \)
- \( l^{(k)} \), the correct ranking of documents for query \( q_k \)
- \( l \in L \setminus l^{(k)} \), a ranked list of documents which differs from \( l^{(k)} \)
- \( \psi(q_k, l) \in \mathbb{R}^m \), the feature vector representation of ranked list \( l \) w.r.t. query \( q_k \)
- \( \Delta(l^{(k)}, l) \), the AP loss of ranked list \( l \) in comparison to the correct ranking \( l^{(k)} \)
- \( w \), the vector of model parameters in rankSVM or SVM-MAP
- \( \xi_{ij} \) and \( \xi_k \), the slack variables in rankSVM and SVM-MAP, respectively.

Answers:
3. Clustering [15 points]

Suggest two (or more) real-world applications (tasks) for which you believe that unsupervised clustering would offer significant help; explain why you believe so. Choose the most appropriate clustering method for each task from agglomerative, top-down, k-means, single pass and EM-based clustering, or any combination of them, based on the properties of the methods and the nature of the tasks. Justify your choices. Creative thoughts and insightful analyses are encouraged.
4. Collaborative Filtering [25 points]

Collaborative filtering systems can use different similarity measures between users and items. A measure allows us to determine how two elements are related to each other; thus a good measure would yield large values for similar users or items, but small values for pairs that are very different. This question asks you how to compute commonly used measures and understand when they should be used.

The matrix below presents the ratings given by 3 users (W, X, Y) for 7 items (a-g) on a 1-5 scale:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Part I: Measure similarity for vectors with multi-level rates [6 points]

1. Using the dot-product as a measure, compute the similarity between the pairs of the users: [2 points]
   - W and X:
   - X and Y:
   - W and Y:

2. Adjust the rates by subtracting the average rating given by its user (do this only on non-empty cells) and compute the L2 norm for each user profile (row vector): [2 points]

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>L2 norm squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Compute the PPC using the matrix that you created in part 2 (you can write your answer as a fraction): [2 points]
   - W and X:
   - X and Y:
   - W and Y:
Part II: Measure similarity for Boolean vectors [10 points]

1. In many applications, the matrix elements are Boolean valued (i.e. 0 or 1). Could you think of applications in which one would need to use a Boolean valued matrix? Please list 2 applications and describe briefly what the matrix represents. [4 points]

2. Convert the matrix given in part 1 into a binary matrix. Replace the non-zero entries with 1. [2 points]

\[
\begin{array}{cccccccc}
 & a & b & c & d & e & f & g \\
W & & & & & & & \\
X & & & & & & & \\
Y & & & & & & & \\
\end{array}
\]

3. Compute the similarities on this matrix using the dot-product of: [2 points]
   - W and X:
   - X and Y:
   - W and Y:

4. Compute the Jaccard similarity on the binary-valued matrix for the following pairs of users. You can leave your results as fractions: [2 points]
   - W and X:
   - X and Y:
   - W and Y:

**Definition**: The Jaccard similarity is a set-based measure. It is defined as follows: 
\[
J(A,B) = \frac{|A \cap B|}{|A \cup B|}
\]
where A and B are two sets. That is: the Jaccard similarity between two sets is the size of the intersection divided by the size of the union of both sets. For example, for A={1, 3, 5} and B={2, 3, 4, 5} \(J(A,B) = 2/5\). For applying Jaccard similarity to boolean vectors, we treat each vector as a set. For example: if we have the vectors A = [0, 1, 1, 1] and B=[1, 0, 1, 1], we convert them to sets by taking the indexes of the non-zero elements (the first position of the vector has index 1) as A= \{2, 3, 4\}, B = \{1, 3, 4\}. Then we can compute the Jaccard similarity as: \(J(A, B) = |\{3, 4\}| / |\{1, 2, 3, 4\}| = 1/2\).
Part III: Collaborative Filtering in practice [9 points]

As you know, collaborative filtering is used in many kinds of recommendation systems. However, those systems are usually effective only when they have collected enough information about a user. A common problem appears when the system has to recommend items to a new user who typically does not have many rated any items.

Recommend some approaches (at least two) that you believe to be effective for addressing the data-sparse issue with new users in collaborative filtering. What kind of information would you leverage in those approaches?
5. **Federated Search** [20 points]

Suppose that you are developing a federated search service that includes fifty different information services (e.g., search engines, stock quotes, flight status, package tracking, ...).

a. What algorithm or approach would you use for resource selection, and why? [8 points]

b. One of the information services is a news search engine that contains content obtained from the New York Times. News is updated constantly, thus the material in the news search engine is changing constantly. The New York Times does not want your news search engine to compete with theirs, so your news search engine can only be accessed through your federated search system; your customers cannot access it directly. How does your resource selection algorithm know which queries to direct to the news search engine? Which of its component pieces is the most important for handling news queries? Discuss the strengths and weaknesses of your solution (i.e., what it will handle well, and won’t it won’t handle well). [12 points]