CS47300 Fall 2017 Midterm 2 and solutions, November 10, 2017
Prof. Chris Clifton

Turn Off Your Cell Phone. Use of any electronic device during the test is prohibited. As previously noted, you are allowed notes: Up to four sheets of 8.5x11 or A4 paper, single-sided (or two sheets double-sided).

Time will be tight. If you spend more than the recommended time on any question, go on to the next one. If you can’t answer it in the recommended time, you are either giving too much detail or the question is material you don’t know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

Note: It is okay to abbreviate in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

In all cases, it is important that you give some idea of how you derived the answer, not simply give an answer. Setting up the derivation correctly, even if you don’t carry out the calculations to get the final answer, is good for nearly full credit.

1 Text Categorization (5 minutes, 4 points)

Given the following six documents, divided into two categories, do you think k-Nearest Neighbor or Support Vector Machine would be a more effective classifier? Explain why, giving specifics. It is okay to assume k=1 for k-NN.

<table>
<thead>
<tr>
<th>Category 1</th>
<th>Category 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1: cricket bat play</td>
<td>D4: bat eat cricket</td>
</tr>
<tr>
<td>D2: play club sports</td>
<td>D5: eat sports club</td>
</tr>
<tr>
<td>D3: hot play dog</td>
<td>D6: eat hot dog</td>
</tr>
</tbody>
</table>

(Note: The documents have been stemmed and had stopwords removed, they would make more sense in the original... But you don’t need to worry about that.)

The data is clearly linearly separable (along the “play” dimension, for example, presence or absence of play divides the categories.) k-NN seems unlikely to work well, for example, the nearest neighbor of D1 is D4, which is in a different category (for the more ML-minded, if we did a leave-one-out cross validation, 1-NN would perform miserably.) Since SVM is a linear separator, it would be my choice.

Scoring: 1 for ”SVM”, 1 for noting SVM is linear separator, 1 for k-nn isn’t linear separator, 1 for showing data linearly separable, 1 for showing nearest neighbor often wrong.

2 Text Categorization (5 minutes, 3 points)

For a text categorization problem where all categories are the same size, is it better to use micro-F1 score or macro-F1 score? Explain your answer, the more formal/mathematical your explanation, the better.

\[
\text{macro-F}_1 = \frac{1}{\#\text{categories}} \sum_{c \in \text{categories}} 2 \cdot \frac{\text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}
\]
\[
\text{micro-F}_1 = 2 \ast \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Note that the key distinction is that macro-F1 averages over categories, so a small category contributes equally to the score with a large category. With micro-F1, each document is treated equally, so small categories contribute very little. Since the categories are the same size, this distinction essentially disappears.
Score: 1 for "doesn’t matter", 1 for noting that the key distinction is that macro-F1 heavily weights small categories while micro-F1 essentially ignores them, 1 for some demonstration of why (e.g., formula in explanation).

3 PageRank (5 minutes, 5 points)

Given the graph below (both graph and adjacency matrix shown), what would go wrong if we apply the PageRank algorithm on it? What could we do to fix this problem?

```
[0 0 0 1
 1 0 0 1
 1 1 0 1
 0 0 0 0]
```

The problem is that the definition of pagerank means that a node with no incoming links ends up with score 0. We can see this using the “random surfer” interpretation of Pagerank, where a random surfer never gets to 2 (unless they happen to start there), and up “stuck” at 3 (so all time is spent there.)

To fix this, we can add a link from every node to every other node, but with a low weight. In the matrix view, we do this by adding a small $\beta$ to each node, so there are no non-zero entries.

Scoring: 1 for nothing that the problem is with node 3 or 2, 2 for noting that the problem is with no incoming or no outgoing links leading to 0 or all weight flowing to sink. For fixing, 1 for some reasonable idea, 2 for correctly demonstrating understanding of adding links to/from everywhere or adding teleport to algorithm. 1 for a formal statement or mathematical demonstration of either.

4 PageRank vs. HITS (5 minutes, 5 points)

Which would you say is better, the PageRank algorithm or the HITS algorithm? Explain your answer, try to give at least two reasons why one or the other is better.

HITS. (1pt)

PageRank algorithm compared to HITS is more efficient (1 pt), since it computes a single measure of quality for a page at crawl time. This measure is then combined with traditional information retrieval score at query time. (1 pt) Therefore, it gives much greater efficiency than HITS.

However, PageRank does not attempt to capture the distinction between hubs and authorities, (1 pt) while HITS emphasizes mutual reinforcement between hub and authority web pages. (1 pt)

5 Web Crawling (5 minutes, 3 points)

Assume that we have two different approaches to crawl. Approach A is to get documents through random IP addresses, and Approach B uses random walks.
Assume we are doing a crawl within a company, where we have access to the Domain Name Server listing all IP addresses. Which approach do you think is better? Explain why.

The problem with Random IP is that we spend a lot of time chasing IPs that don’t exist or don’t have a web server. Since we have the DNS information, we can randomly choose from the IPs that actually exist, avoiding this problem. Random walks may miss disconnected networks (anything that doesn’t have a link from another site in the company), so we may never get complete coverage.

Scoring: 1 for “Random IP”, 1 for “Random walks may miss disconnected networks”, 1 for “Random IP can be chosen to always get a valid machine”

6 Web Crawling (10 minutes, 5 points)

A. We sample URLs uniformly at random from search engine A and check if they are present in search engine B. Its result is that 30% of A’s pages are found in B, and 50% of B’s pages are found in A. Determine A’s size relative to B.

\[ A \cap B = 0.3A = 0.5B, \text{ so } A = \frac{5}{3}B. \]

Scoring: 1 for \( A = \frac{5}{3}B \) or \( B = \frac{3}{5}A \), 1 for derivation or graph or explanation why.

B. Explain how you can use your results from 1 to estimate the total size of the web. List what other information you need, and what assumptions you make.

If we assume that B is a random sample of the internet, then we can state that for a randomly chosen web page (i.e., something that appears in B), \( P(x \in A) = 0.5 \). Therefore, A covers half the web, if we know the size of A, then the web is twice that size.

1 for “size of A or B”, 1 for derivation/explanation, 1 for \( 2 \cdot A \) or \( \frac{6}{5}B \), 1 for “random sample”

7 Text Clustering (10 minutes, 5 points)

Given the following documents:

D1: There is crop loss in Indiana due to the summer heat

D2: Hot summer affects crop yield in Indiana

D3: Scientists discover a new drug that can delay aging

D4: Scientists discover a pill that can delay aging effects to keep you young

D5: Scientists say outdoor sports help improve minds at young age

D6: Outdoor sports help improve attention spans of young minds

After stemming and removing stop words, a term-document incidence matrix is obtained, and Inter-document Euclidean distances are calculated. The distance matrix for the documents is provided below:
Perform k-means clustering with $k = 2$ (with initial cluster centers $D1$ and $D6$) and $k = 3$ (with initial cluster centers $D1$, $D3$ and $D5$) for the above documents. Report the average inter-cluster Euclidean distance in each case. Which is a better $k$ for the given corpus and why?

Average inter-cluster Euclidean distance for a cluster is defined as the average of pair-wise Euclidean distances between all documents in the given cluster. The Mean of that quantity over all clusters is the evaluation metric we ask you to report.

Note the typo in the above: We said inter-cluster, but defined it as intra-cluster. As a result, there are a number of answers accepted as correct (either a reasonable calculation of intra-cluster or inter-cluster).

$K = 2$

$D1$: $D1$, $D2$, $D3$ \hspace{1cm} ICD(C1) = \frac{2 + 3.32 + 3.32}{3} = 2.88

$D6$: $D4$, $D5$, $D5$ \hspace{1cm} ICD(C2) = \frac{3.46 + 3.74 + 2}{3} = 3.07

ICD = \frac{2.88 + 3.07}{2} = 2.975

$K = 3$:

$D1$: $D1$, $D2$ \hspace{1cm} ICD(C1) = 2

$D3$: $D3$, $D4$ \hspace{1cm} ICD(C2) = 2.24

$D6$: $D5$, $D6$ \hspace{1cm} ICD(C3) = 2

ICD = \frac{2 + 2.24 + 2}{3} = 2.08

$K=3$ is better.

1 for each proper clustering, 1 each for intra-cluster distances, 1 for good explanation of why they chose one as the best, 1 for noting that intra-cluster distance is maximized for $k = n$ or several other good insights, up to a total of four.